

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年9月24日(24.09.2020)



(10) 国際公開番号

WO 2020/188971 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号: PCT/JP2020/000837

(22) 国際出願日: 2020年1月14日(14.01.2020)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2019-048115 2019年3月15日(15.03.2019) JP

(71) 出願人: 富士フイルム株式会社 (**FUJIFILM CORPORATION**) [JP/JP]; 〒1068620 東京都港区西麻布2丁目2番30号 Tokyo (JP).

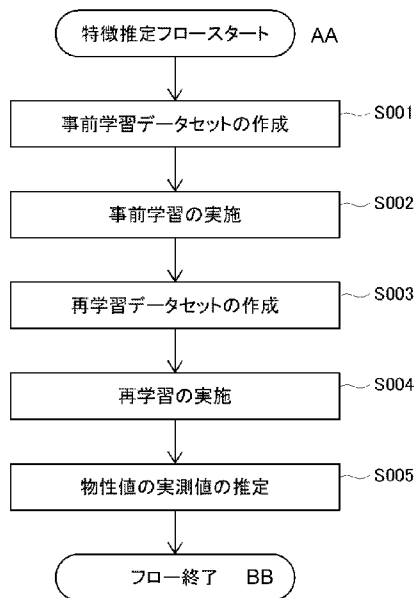
(72) 発明者: 柴田 洋平(**SHIBATA Yohei**); 〒2588577 神奈川県足柄上郡開成町牛島577番地 富士フイルム株式会社内 Kanagawa (JP).

(74) 代理人: 伊東 秀明, 外 (**ITO HIDEAKI et al.**); 〒1010032 東京都千代田区岩本町2丁目3番3号 ザイマックス岩本町ビル6階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,

(54) **Title:** CHARACTERISTICS ESTIMATION METHOD, CHARACTERISTICS ESTIMATION DEVICE, PROGRAM, AND RECORDING MEDIUM

(54) 発明の名称: 特徴推定方法、特徴推定装置、プログラム及び記録媒体



S001 Create pre-learning data sets
S002 Perform pre-learning
S003 Create relearning data sets
S004 Perform relearning
S005 Estimate observed values of physical properties
AA Start of characteristics estimation flow
BB End of flow

(57) **Abstract:** The present invention highly accurately estimates characteristics of an object using machine learning, regardless of the number of characteristics identified under targeted estimation conditions. Machine learning is performed using descriptive information relating to the configuration of an object, identification information about conditions for identifying characteristics, and characteristics of the object. The machine learning includes pre-learning and relearning. The pre-learning creates both an identification information calculation model, which receives identification information

NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

- 一 国際調査報告 (条約第21条(3))

about each first condition and outputs calculated identification information, and a nonlinear estimation result output model, which receives the calculated identification information and either the descriptive information or information obtained from the descriptive information and outputs characteristics estimation results. The relearning updates the parameters in the identification information calculation model on the basis of identification information about a second condition and on the basis of characteristics that have been identified under the second condition, while maintaining constant the parameters in the estimation result output model, and creates an estimation model that estimates, from the descriptive information and the identification information about the second condition, characteristics identified under the second condition.

(57) 要約：機械学習を利用して対象の特徴を推定する際に、推定の狙いとする条件の下で特定された特徴の個数に依らず、高精度に推定する。対象の構成に関する記述情報、特徴を特定する条件の識別情報、及び対象の特徴を用いた機械学習を実施する。機械学習は、事前学習と再学習とを含む。事前学習では、各第一条件の識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、記述情報又は記述情報から得られた情報と演算後識別情報とが入力されて特徴の推定結果を出力する非線形型の推定結果出力モデルと、を構築する。再学習では、推定結果出力モデル中のパラメータを固定したまま識別情報用演算モデル中のパラメータを第二条件の識別情報、及び、第二条件の下で特定された特徴に基づいて更新し、記述情報及び第二条件の識別情報から第二条件の下で特定される特徴を推定する推定モデルを構築する。

明 細 書

発明の名称：

特徴推定方法、特徴推定装置、プログラム及び記録媒体

技術分野

[0001] 本発明は、対象の特徴を推定する特徴推定方法及び特徴推定装置に係り、特に、機械学習を実施して構築した推定モデルによって特徴を推定する特徴推定方法及び特徴推定装置に関する。

また、本発明は、上記の推定モデルを構築する処理をコンピュータに実行させるためのプログラム、及び当該プログラムを記録した、コンピュータに読み取り可能な記録媒体に関する。

背景技術

[0002] 対象の特徴と、その特徴を特定するための条件との関係をニューラルネットワーク等の機械学習によってモデル化し、ある条件の下で特定される対象の特徴を推定する技術は、既に知られている。例えば、ニューラルネットワークを用いて化合物の特性を予測する技術が、特許文献 1 に記載されている。

[0003] 特許文献 1 に記載の技術は、「化学混合物の色以外の特性を予測する方法であって、化学混合物の成分量を含む化学混合物変数と、環境及びプロセス条件と、実測された化学混合物の特性と、を含むデータを収集する工程と、実測された化学混合物の特性に対して化学混合物変数が及ぼす影響を明らかにするニューラルネットワークを作製する工程と、上記のデータを用いてニューラルネットワークを教師付き学習する工程と、ニューラルネットワークによって新たな化学混合物の特性を前方予測する工程と、を有する方法」である。この技術によれば、化学混合物変数と化学混合物の特性の実測結果との関係を明らかにするためにニューラルネットワークを学習し、学習されたニューラルネットワークを用いて、新たな化学混合物の特性を予測することができる。この結果、追加的に実験を行うことなく、新たな化学混合物の特

性を予測（推定）することが可能となる。このような技術は、マテリアルズ・インフォマティクスと呼ばれ、材料開発の分野等において注目されている。

先行技術文献

特許文献

[0004] 特許文献1：特表2008-509486号公報

発明の概要

発明が解決しようとする課題

[0005] しかし、特許文献1に記載の技術を利用する場合には、化学混合物の特性の実測データを用いて学習（教師付き学習）する必要がある。ただし、化学混合物の特性の実測データを取得するには実験等を要するため、実測データの取得数が比較的少ない状況が考えられる。そして、少ない実測データを用いて学習を行った場合には、その少ない実測データに特化した予測モデルが構築される結果、「過学習」の問題が生じ得る。

[0006] ここで、化学混合物の特性の実測データは、予測の狙いとする条件（実測条件）の下で特定されたデータ（以下、「ターゲットデータ」とも言う。）である。他方、化学混合物の特性については、第一原理計算等の計算式を用いて計算することが可能であり、そのような計算結果のデータ（以下、「補助データ」とも言う。）をターゲットデータの代わりに利用して学習を行うことも考えられる。

[0007] また、化学混合物の特性を予測する上では、あくまでも、予測の狙いとする条件の下で特定される特性（つまり、特性の実測値）を予測することが求められる。したがって、ニューラルネットワーク等の機械学習を利用して、予測の狙いとする条件の下で特定される特性を予測する場合には、ターゲットデータの量に拘わらず、高精度で予測することが必要となる。

[0008] 本発明は、上記の事情に鑑みてなされたものであり、以下に示す目的を解決することを課題とする。

つまり、本発明は、上記従来技術の問題点を解決し、機械学習を利用して対象の特徴を推定する際に、推定の狙いとする条件の下で特定された特徴の個数に依らず、高精度に推定することが可能な特徴推定方法及び特徴推定装置を提供することを目的とする。

また、本発明は、上記の特徴推定方法をコンピュータによって実現するためのプログラム、及び、当該プログラムを記録した記録媒体を提供することをも目的とする。

課題を解決するための手段

[0009] 上記の目的を達成するために、本発明の特徴推定方法は、コンピュータにより、対象の特徴を推定する特徴推定方法であって、コンピュータが、対象の構成に関する記述情報、特徴を特定する条件に対して設定された識別情報、及び条件の下で特定された特徴を用いた機械学習を実施して、記述情報及び識別情報から特徴を推定する推定モデルを構築し、条件は、複数の第一条件と、複数の第一条件とは異なる第二条件と、を含み、機械学習は、記述情報、複数の第一条件の各々の識別情報、及び、複数の第一条件の各々の下で特定された特徴を用いた事前学習と、第二条件の識別情報、及び、第二条件の下で特定された特徴を用いた再学習と、を含み、事前学習では、コンピュータが、複数の第一条件の各々の識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、記述情報又は記述情報から得られた情報と演算後識別情報とが入力されて特徴の推定結果を出力する非線形型の推定結果出力モデルと、を構築し、再学習では、コンピュータが、推定結果出力モデル中のパラメータを固定したまま識別情報用演算モデル中のパラメータを第二条件の識別情報、及び、第二条件の下で特定された特徴に基づいて更新し、記述情報及び第二条件の識別情報から第二条件の下で特定される特徴を推定する推定モデルを構築することを特徴とする。

[0010] 以上のように構成された本発明の特徴推定方法によれば、推定の狙いとする条件（第二条件）の下で特定された特徴の個数に拘わらず、当該条件（第二条件）の下で特定される特徴を高精度にて推定することが可能となる。

[0011] また、事前学習では、コンピュータが、記述情報が入力されて演算後記述情報を出力する記述情報用演算モデルと、識別情報用演算モデルと、演算後記述情報と演算後識別情報とが入力されて特徴の推定結果を出力する推定結果出力モデルと、を構築し、再学習では、コンピュータが、記述情報用演算モデル中のパラメータ、及び、推定結果出力モデル中のパラメータを固定したまま識別情報用演算モデル中のパラメータを第二条件の識別情報、及び、第二条件の下で特定された特徴に基づいて更新し、記述情報及び第二条件の識別情報から第二条件の下で特定される特徴を推定する推定モデルを構築すると、好適である。

上記の構成であれば、事前学習で推定モデル全体のパラメータを設定し、再学習で、推定モデルの一部である識別情報用演算モデル中のパラメータのみを更新することになる。

[0012] また、演算後記述情報は、多次元の第一ベクトルであり、演算後識別情報は、第二ベクトルであり、推定結果出力モデルは、第一ベクトル及び第二ベクトルから特徴の推定結果を出力するモデルであると、より好適である。

[0013] また、第二ベクトルの次元数は、記述情報用演算モデル中のパラメータの数よりも小さく、且つ、推定結果出力モデル中のパラメータの数よりも小さいと、より好適である。

上記の構成であれば、推定の狙いとする条件（第二条件）の下で特定された特徴の個数が少ない場合であっても、二段階の学習によって、推定精度がより高い推定モデルを構築することが可能である。

[0014] また、第二ベクトルの次元数は、再学習で用いられる第二条件の下で特定された特徴の数よりも小さいと、より一層好適である。

上記の構成であれば、機械学習において過学習の発生が抑えられる。

[0015] また、識別情報用演算モデルは、条件毎に設定された固有次元ベクトルによって構成される行列を含む演算モデルであり、再学習では、識別情報用演算モデル中のパラメータとして、行列中の成分値を更新すると、好適である。

[0016] また、機械学習において、コンピュータは、ニューラルネットワークにより事前学習及び再学習を実施すると、好適である。

上記の構成であれば、事前学習にて、推定される特徴の正解を表現することが可能な関数（以下、「正解関数」とも言う。）が属する関数族を適切に見つけ出し、再学習にて、正解関数を近似する推定モデルを特定することが可能である。

[0017] また、推定モデルは、対象の特徴としての化合物の物性値を推定するモデルであり、記述情報は、化合物の構造を線形表記法に則って記述した情報であり、条件は、化合物の物性値を特定するために設定された条件であってもよい。

上記の構成であれば、化合物の物性値を推定することが可能な推定モデルを構築することができ、このモデルを利用することで、未知である化合物の物性値を高精度に推定することが可能となる。

[0018] また、第一条件は、物性値を計算するために設定された計算条件であり、第二条件は、物性値を実測するために設定された実測条件であってもよい。

上記の構成であれば、未測定である化合物の物性値を高精度に推定することが可能となる。

[0019] また、前述した課題を解決するために、本発明の特徴推定装置は、対象の特徴を推定する特徴推定装置であって、対象の構成に関する記述情報、特徴を特定する条件に対して設定された識別情報、及び条件の下で特定された特徴を用いた機械学習を実施して、記述情報及び識別情報から特徴を推定する推定モデルを構築するモデル構築部を有し、条件は、複数の第一条件と、複数の第一条件とは異なる第二条件と、を含み、機械学習は、記述情報、複数の第一条件の各々の識別情報、及び、複数の第一条件の各々の下で特定された特徴を用いた事前学習と、第二条件の識別情報、及び、第二条件の下で特定された特徴を用いた再学習と、を含み、事前学習では、モデル構築部が、複数の第一条件の各々の識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、記述情報又は記述情報から得られた情報と演算後識

別情報とが入力されて特徴の推定結果を出力する非線形型の推定値出力モデルと、を構築し、再学習では、モデル構築部が、推定値出力モデル中のパラメータを固定したまま識別情報用演算モデル中のパラメータを第二条件の識別情報、及び、第二条件の下で特定された特徴に基づいて更新し、記述情報及び第二条件の識別情報から第二条件の下で特定される特徴を推定する推定モデルを構築することを特徴とする。

[0020] また、本発明は、上述した本発明の特徴推定方法における機械学習を実施して推定モデルを構築する処理をコンピュータに実行させるためのプログラムを提供する。

また、本発明は、上述した本発明の特徴推定方法における機械学習を実施して推定モデルを構築する処理をコンピュータに実行させるためのプログラムが記録された、コンピュータに読み取り可能な記録媒体を提供する。

[0021] さらに、本発明は、プロセッサを備える特徴推定装置であって、プロセッサが、対象の構成に関する記述情報、特徴を特定する条件に対して設定された識別情報、及び条件の下で特定された特徴を用いた機械学習を実施して、記述情報及び識別情報から特徴を推定する推定モデルを構築し、条件は、複数の第一条件と、複数の第一条件とは異なる第二条件と、を含み、機械学習は、記述情報、複数の第一条件の各々の識別情報、及び、複数の第一条件の各々の下で特定された特徴を用いた事前学習と、第二条件の識別情報、及び、第二条件の下で特定された特徴を用いた再学習と、を含み、事前学習では、プロセッサが、複数の第一条件の各々の識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、記述情報又は記述情報から得られた情報と演算後識別情報とが入力されて特徴の推定結果を出力する非線形型の推定結果出力モデルと、を構築し、再学習では、プロセッサが、推定結果出力モデル中のパラメータを固定したまま識別情報用演算モデル中のパラメータを第二条件の識別情報、及び、第二条件の下で特定された特徴に基づいて更新し、記述情報及び第二条件の識別情報から第二条件の下で特定される特徴を推定する推定モデルを構築するように構成された特徴推定装置を提供

する。

発明の効果

[0022] 本発明によれば、機械学習を利用して対象の特徴を推定する際に、推定の狙いとする条件の下で特定された特徴の個数に依らず、高精度に推定することが可能な特徴推定方法及び特徴推定装置が提供される。

また、本発明によれば、上述した本発明の特徴推定方法をコンピュータによって実現するためのプログラム、及び当該プログラムを記録した記録媒体をも提供することが可能である。

図面の簡単な説明

[0023] [図1]化合物の物性値を条件毎に示す表である。

[図2]転移学習の問題点についての説明図である。

[図3]機械学習によって導出される関数族についての説明図である。

[図4]本発明の機械学習による効果を示す図である。

[図5]本発明の一実施形態に係る推定モデルの構成を示す図である。

[図6]識別情報用演算モデルについての説明図である。

[図7]本発明の一実施形態に係る特徴推定装置のハードウェア構成を模式的に示す図である。

[図8]本発明の一実施形態に係る特徴推定装置の機能部を示すブロック図である。

[図9]特徴推定フローの流れを示す図である。

[図10]事前学習にて出力された第二ベクトルを可視化した図である。

[図11]再学習にて追加された第二ベクトルを可視化した図である。

[図12]検証実験で用いた物性値の計算データの一部を示す図である。

[図13]検証実験の結果を示す図である。

発明を実施するための形態

[0024] 本発明の一実施形態（以下、「本実施形態」と言う。）に係る特徴推定方法、特徴推定装置、プログラム及び記録媒体について、添付の図面を参照しながら、以下に詳細に説明する。

なお、以下に説明する実施形態は、本発明を分かり易く説明する目的で挙げた一例にすぎず、本発明を限定するものではない。すなわち、本発明は、下記の実施形態に限られず、本発明の主旨を逸脱しない範囲において種々の改良又は変更され得る。また、当然ながら、本発明には、その等価物が含まれる。

[0025] <本実施形態の概要>

本実施形態では、機械学習を利用して対象の特徴を推定する。ここで、「対象」とは、物質、人又は動植物、構造物、若しくはその他の有体物であってもよく、あるいは、音、光、香気、若しくはその他の無体物であってもよい。

また、「特徴」とは、対象が有する性質又は特性、対象に対する評価、対象が属する属性、若しくは対象から認識される情報（例えば、対象が音声である場合には、その音声が意味する語句等）等である。なお、本実施形態では、化合物の物性値を「対象の特徴」として推定することとする。

[0026] 本実施形態では、互いに異なる条件にて取得された物性値を示すデータを教師データとして機械学習を実施し、未知の物性値を推定する数理モデル（以下、「推定モデル」と言う。）を構築する。ここで、条件とは、物性値を特定するために設定された条件であり、本実施形態では、条件が互いに異なる2種類の条件（本発明の「第一条件」及び「第二条件」）に大別される。

[0027] より詳しく説明すると、本実施形態では、既知のターゲットデータ及び既知の補助データを用いた機械学習を実施して、未知のターゲットデータを推定する推定モデルを構築する。本実施形態に係るターゲットデータは、推定の狙いとする条件の下で特定される物性値であり、具体的には、物性値の実測値である。なお、既知のターゲットデータは、実験を行って取得されるため、その取得数は、比較的少ない。

[0028] ここで、既知のターゲットデータは、実測条件（図1の「実測A」）の下で特定された物性値の実測値であり、図1の最上段に示された値である。図1は、複数種類の化合物の各々の物性値を条件毎に特定した結果を示す表で

ある。なお、図1には、化合物の物性値として、有機化合物の光の吸収エネルギー量が示されている。

[0029] 図1に示すケースでは、既知のターゲットデータ（すなわち、物性値の実測値）は、数種類の化合物についてのみ得られている。一方、化合物の中には、物性値の実測値を欠くものが存在し、この欠落した実測値が未知のターゲットデータに相当する。なお、以下では、既知のターゲットデータをまとめたデータ群を「ターゲットデータソース」と呼ぶこととする。また、未知のターゲットデータについて仮に実験を行った場合に補完される物性値（すなわち、実測値）を「正解データ」と呼ぶこととする。

[0030] また、本実施形態に係る補助データは、ターゲットデータが得られる条件以外の条件の下で特定される物性値であり、具体的には、物性値を第一原理計算の計算式等によって計算した際の計算値である。化学分野では、同一化合物の吸収波長をシミュレーションによって計算する計算手法が多数あり、それぞれの手法にて異なった計算結果が得られる。

[0031] より具体的に説明すると、第一原理計算を用いる場合には、汎関数、溶媒効果及び基底関数等の組み合わせによって計算手法が指定されることになる。ここで、どの計算手法によって得られた計算結果が真の物性値（つまり、実測値）を良好に再現しているかについては、計算対象の化合物及び実験環境等によって変わってくる。つまり、実測値の再現精度の観点から評価した上記の計算手法の優劣は、一般的に、事前に把握することが困難であり、実際の優劣評価は、各計算手法による計算結果と実測値とを対比することで行われる。

[0032] 以上の理由から、物性値の計算結果を示す補助データは、複数得られ、具体的には、化合物の種類数と計算手法の種類との積に相当する数のデータが得られる。すなわち、図1から分かるように、既知の補助データは、既知のターゲットデータに比して圧倒的に多く、それぞれの種類の計算手法（具体的には、図1の「計算B」～「計算Z」）について化合物別に得られる。

なお、以下では、数千～数万件の補助データが入手された状況を想定して

説明することとする。また、以下では、既知の補助データをまとめたデータ群を「補助データソース」と呼ぶこととする。

[0033] また、補助データの取得条件（図1の「計算B」～「計算Z」）は、物性値を第一原理計算によって計算するために設定された計算条件であり、本発明の「第一条件」として複数設定されている。これに対して、既知のターゲットデータの取得条件（図1の「実測A」）は、物性値を実測するために設定された実測条件（換言すると、実験条件）であり、本発明の「第二条件」として少なくとも一つ設定されている。本実施形態では、ターゲットデータの取得条件を1つのみとするが、当然ながら、ターゲットデータの取得条件を複数設定してもよい。

[0034] 本実施形態では、上述したように、既知のターゲットデータ及び既知の補助データを用いて機械学習を実施し、未知のターゲットデータ（未測定である化合物の物性値）を推定する推定モデルを構築する。本実施形態では、機械学習が事前学習と再学習からなり、それぞれの学習がニューラルネットワークによって実施される。ニューラルネットワークは、入力層及び出力層の少なくとも2層の処理要素を含む。処理要素は、相互に結合され、処理要素間に所定の結合重み付けをもった所定のパターン（関数）とされている。

[0035] そして、推定モデルを構成するニューラルネットワークは、入力の変化に対する化合物の物性値の応答をシミュレートするように予め学習させられる。この学習により、処理要素間の結合重み付けが、化合物及びその物性値の取得条件に関する情報（入力）と当該物性値（出力）との間の関係を反映したものになる。この結果、上記のニューラルネットワークにより、化合物の種類及び物性値の取得条件のそれぞれの組み合わせに対して、当該組み合わせに対応する物性値を推定することができる。なお、機械学習は、ニューラルネットワークに限定される訳ではなく、多項式回帰、若しくは一般化線形モデル等であってもよい。ただし、後述する識別情報用演算モデルの更新を行うことから、入力データに対して微分が可能な学習である必要がある。

[0036] 機械学習について概説すると、本実施形態に係る機械学習では、前述のよ

うに、先ず事前学習を実施し、その後に再学習を実施する。事前学習では、既知の補助データを含む学習データセット（以下、「事前学習データセット」と言う。）を用いて推定モデルを構築する。ここで、事前学習データセットは、補助データが得られた化合物の構成に関する記述情報、及び、補助データの取得条件の識別情報、並びに、補助データが示す物性値の計算値（換言すると、補助データの取得条件の下で特定された物性値）を1組とするデータセットであり、具体的には図1のデータマトリクス中、2段目以降の各段のデータから作成される。なお、記述情報及び識別情報については、後の項で詳しく説明することとする。

[0037] そして、本実施形態では、既知の補助データと同数（すなわち、数千～数万個）の事前学習データセットが作成され、作成した事前学習データすべてを用いて事前学習を実施する。そして、事前学習により推定モデルが構築されるが、換言すると、推定モデル中のパラメータ（具体的には、後述のパラメータ θ 、 ϕ 、 ψ ）が設定される。

[0038] 再学習では、既知のターゲットデータを含む学習データセット（以下、「再学習データセット」と言う。）を用いて推定モデルの一部を再構築する。ここで、再学習データセットは、ターゲットデータが得られた化合物の構成に関する記述情報、及び、ターゲットデータの取得条件の識別情報、並びに、ターゲットデータが示す物性値の実測値（換言すると、ターゲットデータの取得条件の下で特定された物性値）を1組とするデータセットであり、具体的には図1のデータマトリクス中、最上段のデータから作成される。

[0039] ここで、再学習データセットの作成数は、少数（数個～数十個程度）であるため、再学習は、少数の再学習データセットを用いて実施されることになる。換言すると、本実施形態では、再学習に用いられる物性値の実測値（つまり、第二条件の下で特定された特徴）のデータ数が、事前学習に用いられる物性値の計測値（つまり、第一条件の下で特定された特徴）のデータ数よりも圧倒的に少なくなっている。

[0040] そして、再学習により推定モデルの一部が再構築され、換言すると、推定

モデル中の一部のパラメータ（具体的には、後述のパラメータ θ ）が更新される。この結果、未知のターゲットデータ、すなわち未測定である化合物の物性値を推定する推定モデルが構築される。この推定モデルによれば、未測定である化合物の物性値について、精度よく実測値を推定（予測）することができるので、実験を行うことなく実測値の欠落を補完することが可能となる。

[0041] <従来の特徴推定手法について>

本実施形態の詳細を説明するにあたり、その比較例として、化合物の物性値を推定する従来手法について、3つの例を挙げて説明することとする。

[0042] （従来手法1）

従来手法1は、既知のターゲットデータと既知の補助データとの関係を定量化し、その関係を利用して、未知のターゲットデータを予測する回帰分析法である。具体的に説明すると、従来手法1では、例えば、図1に示すデータマトリクス中、最上段のターゲットデータ（すなわち、実測値）と2段目の補助データ（すなわち、計算Bによる計算結果）とに着目し、両方のデータが揃っている化合物（図1では、化合物1、化合物3及び化合物N）を選び、それぞれの化合物のターゲットデータ及び補助データの間で回帰分析を実施する。これにより、回帰モデル（具体的には、回帰式）が分析結果として得られる。この回帰モデルに、実測値が未知である化合物の計算結果を入力すると、実測値の推定値が得られる。

[0043] 以上に説明した従来手法1では、多数取得された補助データのうち、ごく一部のデータのみを分析に用いるので、分析結果が取得した補助データのすべてを反映したものとなっていない虞がある。また、回帰モデルが示す関数（具体的には、回帰式）が、ターゲットデータと補助データとの関係を的確に表現できるように設定されなければ、推定精度が低くなるという問題がある。

[0044] （従来手法2）

従来手法2は、化合物の構造と共に、化合物の物性値の取得条件を記述情

報として表現し、それぞれの記述情報を入力とする推定モデルを構築する手法である。具体的に説明すると、各化合物の構成に関する記述情報と、様々な物性値の取得条件（具体的には、実測条件及び計算条件）を表す記述情報と、各取得条件の下で特定した各化合物の物性値と、を1組とする学習データセットを作成し、その学習データセットを用いた機械学習を実施して、推定モデルを構築する。この推定モデルによれば、ある化合物の構成に関する記述情報と、実測条件に関する記述情報と、を入力すると、ある化合物の物性値の実測値を推定（予測）することができる。

[0045] 以上に説明した従来手法2では、物性値の取得条件の記述情報を人手で設定するため、記述情報が適切に設定されない虞がある。より具体的に説明すると、取得条件に関わる事項が複数存在する場合、その中から、物性値に影響を及ぼす重要な事項を記述情報として選出する必要がある。仮に、この重要事項が選出されなければ、物性値の推定精度が低下する虞がある。また、重要事項が選出されたとしても、その内容が記述情報として数値化できないものであるならば、推定モデルに組み込めず、結果として推定モデルを適切に構築することができない虞がある。

[0046] （従来手法3）

従来手法3は、既知の補助データを用いて事前学習を実施した後に、少数の既知のターゲットデータを用いて再学習を実施することで推定モデルを構築する手法である。かかる点において、従来手法3は、本実施形態（すなわち、本発明の特徴推定方法）と共通する。ただし、従来手法3は、1種類の条件（具体的には、1種類の計算条件）の下で得られた補助データを用いて事前学習を実施する点で本実施形態と異なる。

[0047] 従来手法3について詳しく説明すると、従来手法3は、いわゆる転移学習である。転移学習では、まず、既知の補助データ（例えば、図1中の2段目に示す計算Bによる計算結果）と、そのデータが得られた化合物の構成に関する記述情報とを1組とする学習データセットを作成し、この学習データセットを用い、ニューラルネットワークによる事前学習を実施する。事前学習

により、物性値の計算結果を推定するニューラルネットワークの推定モデルが構築され、換言すると、推定モデル中の各パラメータが設定される。

[0048] その後、ターゲットデータ（すなわち、物性値の実測値）と、そのデータが得られた化合物の構成に関する記述情報とを1組とする学習データセットを作成し、この学習データセットを用いて、ニューラルネットワークによる再学習を実施する。再学習により、事前学習により構築されたニューラルネットワーク推定モデルが、物性値の実測値を推定するモデルとして再構築され、換言すると、推定モデル中のパラメータの一部が更新される。

[0049] 以上に説明した従来手法3としての転移学習では、前述したように、多数の補助データのうち、1種類の取得条件の下で得られた補助データのみを事前学習に用いる。その場合には、未知のターゲットデータの正解データ（分かり易くは、未測定である化合物の物性値の実測値）を表現することが可能な正解関数について、当該正解関数が属する関数族を事前学習によって導き出すことが困難である。

[0050] また、転移学習では、再学習の際に更新されるパラメータ（再学習パラメータ）が、ニューラルネットワークによって構成される推定モデルのうち、出力層又は出力層付近の層に存在する。そのため、再学習パラメータの数が少ない場合には、再学習後の推定モデルによって表現される関数が正解関数から外れてしまう虞がある。この問題について詳しく説明すると、一般に、ターゲットデータソースのデータ数が非常に少ない場合には、再学習におけるターゲットデータへの過学習を避けるために、再学習パラメータの数を極力小さく設定する必要がある。ここで、再学習パラメータの数を1と設定したケースを想定する。なお、このようなケースは、例えば、ターゲットデータが数個しか得られていない場合等において現実的に起こり得るケースである。

[0051] 転移学習において再学習パラメータが1である場合、転移学習の再学習は、ベースライン補正と等価となる。ここで、図2に示す具体例を参照しながら具体的に説明すると、事前学習にて構築された推定モデルによって表現さ

れる関数（図2中の関数 f_p ）が、例えば下記式（a1）によって示される関数であるとする。

$$f_p : y = Ax^2 + \theta_1 \quad (a1)$$

上式において、定数 A 及び θ_1 は、事前学習によって設定されるパラメータに該当する。

そして、転移学習において再学習パラメータが1であるときに、少数のターゲットデータを用いて再学習すると、再学習後の推定モデルによって表現される関数（図2中の関数 f_q ）が下記式（a2）によって示される関数となる。つまり、再学習パラメータが1であるときの再学習は、関数 f_p のパラメータ θ_1 を θ_2 にベースライン補正するものとなる。

$$f_q : y = Ax^2 + \theta_2 \quad (a2)$$

[0052] 以上のように、転移学習の場合、再学習パラメータが1である場合に再学習を実施すると、図2に示すように、推定モデル中のパラメータをベースライン補正にて更新する。ただし、事前学習にて構築される推定モデルが示す関数 f_p と正解関数（図2中の関数 f_r ）とが、上述のベースライン補正だけでは説明することができない複雑な関係にある場合、上述した再学習後の推定モデルでは、未知のターゲットデータを推定する際の精度が低くなってしまう。

[0053] <本実施形態に係る推定モデルについて>

本発明者らは、鋭意検討により、上述した従来手法が抱える問題を解決する推定モデルとして、本実施形態に係る推定モデルを完成させた。本実施形態に係る推定モデルについて説明すると、当該推定モデルは、化合物の構成に関する記述情報及びデータ取得条件の識別情報から物性値を推定するモデルである。本実施形態において、推定モデルは、2つのフェーズの機械学習、具体的には前述した事前学習及び再学習によって構築される。

[0054] 事前学習では、多数ある既知の補助データをすべて用いて、ニューラルネットワークによる学習を実施し、推定モデルを構築する。ここで、事前学習で構築される推定モデルは、正解関数が属する関数族（以下、関数族 F ）を

反映している。つまり、事前学習では、図3に示すように、多数の補助データの各々が示す化合物の物性値の計算値に基づき、関数全体の中から関数族Fを導出する。要するに、事前学習は、正解関数を大まかに捉える上で当該正解関数が属する関数族Fを導出することを目的として実施される。

[0055] 再学習では、事前学習にて構築された推定モデル中の一部のパラメータを、少数の既知のターゲットデータに基づいて更新する。これにより、関数族Fの中から、未知のターゲットデータを推定する関数が導出される。つまり、本実施形態では、正解関数が属する関数族Fを事前学習によって導出しておくことにより、その後の再学習に用いられるターゲットデータの数が少数であっても、当該ターゲットデータが示す物性値の実測値を適切に表現し得る複雑な関数を得ることができる。

[0056] ここで、推定モデルのパラメータについて説明すると、本実施形態ではニューラルネットワークによる機械学習が実施されるが、本実施形態では、ニューラルネットワークにおいて膨大な数のパラメータが用いられる。また、ニューラルネットワークによる機械学習によれば、実質的に任意の関数を表現することが可能である。

[0057] また、本実施形態では、ニューラルネットワークによる機械学習のパラメータを2つのグループ（以下、パラメータP、Q）に大別する。殆どのパラメータは、パラメータPに属することとし、極少数のパラメータがパラメータQに属することとする。

[0058] また、ニューラルネットワークによる機械学習によって推定モデルを構築したとき、当該推定モデルによる物性値の推定値 $f(i)$ は、下記式(a3)から求められる。

$$f(i) = (i, P, Q) \quad (a3)$$

上式(a3)において、 i は、推定モデルへの入力情報であり、具体的には、化合物の構成に関する記述情報が該当する。

[0059] また、上式(a3)においてパラメータPを固定すると、物性値の推定値 $f(i)$ は、パラメータQを変数とする関数族とみなすことができる。

- [0060] ここで、各補助データが示す物性値の計算値を表現する関数と、各ターゲットデータが示す物性値の実測値を表現する関数とが、いずれも同一の関数族Fに属する一方で、それぞれの関数間でパラメータが異なると仮定する。より詳しく説明すると、補助データとターゲットデータとの間では、データ取得条件が異なっているが、パラメータPは、その条件に拘わらず共通であると仮定する。他方、パラメータQについては、データ取得条件に応じて変化するものと仮定する。
- [0061] そして、本実施形態では、ニューラルネットワークによる機械学習を実施することで、上述のパラメータを設定する。具体的に説明すると、先ず、多数ある既知の補助データ（物性値の計算値）を用いて事前学習を実施することで、パラメータPを設定する。また、事前学習のフェーズでは、各補助データが示す物性値の計算値を表現するのに必要なパラメータQが設定される。これにより、各補助データが示す物性値の計算値を表現する複数の関数、すなわち、関数族Fを導出することができる。
- [0062] 事前学習の実施後には、少数の既知のターゲットデータ（物性値の実測値）を用いて再学習を実施する。このとき、パラメータPを固定し、ニューラルネットワークによる再学習を実施する。再学習の実施により、パラメータQは、ターゲットデータが示す物性値の実測値を推定するのに適した値へと更新される。なお、パラメータQの数が十分に小さいことから、少数のターゲットデータであっても、パラメータQを適切に設定することが可能である。
- [0063] 以上のような手順により、パラメータP、Qが設定される結果、未知のターゲットデータ（つまり、未測定である化合物の物性値の実測値）を推定することが可能な推定モデルが構築される。そして、推定モデルを用いれば、上述した従来手法の問題を解決し、未知のターゲットデータを精度よく推定することができる。より詳しく説明すると、本実施形態は、既知の補助データ及び既知のターゲットデータをすべて用いる点で、補助データの一部のみを用いる従来手法1（回帰分析）と異なる。また、従来手法1では、ターゲ

ットデータと補助データとの関係を的確に表現するために回帰式の形を予め把握しておく必要がある。これに対して、本実施形態では、ニューラルネットワークによる機械学習が実施されることで、関数族 F が自動的に抽出されるため、関数の式の形を予め把握しておくことを要しない。かかる点において、本実施形態は、従来手法 1 よりも有利である。

[0064] また、本実施形態では、従来手法 2 のように、物性値の取得条件に関する記述情報を人手で設定する必要がなく、具体的には、取得条件の識別情報（具体的には $i d$ 情報）を用いて機械学習を実施する。このため、本実施形態では、取得条件の記述情報が適切に設定されないために物性値の推定精度が低下する事態を回避し、また、物性値の取得条件のうち、物性値の推定精度に影響を及ぼす事項を自動的に抽出することができる。かかる点において、本実施形態は、従来手法 2 よりも有利である。

[0065] また、従来手法 3（転移学習）では、既知の補助データのうち、一種類の条件の下で得られた補助データのみを用いて事前学習を実施するのに対し、本実施形態では、前述したように、多数ある既知の補助データのすべてを用いて事前学習を実施する。これにより、本実施形態では、正解関数が属する関数族 F を良好に推定することが可能である。

[0066] また、転移学習では、再学習パラメータの数が 1 となると、再学習がベースライン補正（単なる定数の加算）となるため、未知のターゲットデータの推定精度が低下する虞がある。これは、転移学習では、再学習パラメータが推定モデルの出力側に存在するためである。これに対して、本実施形態では、後述するように再学習パラメータが推定モデルの入力側に存在する。このため、本実施形態では、ターゲットデータが示す物性値の実測値を表現する関数を、再学習パラメータの数に相当する自由度の範囲内で適切に表現することができる。

[0067] 上記の内容に関して、図 4 を参照しながら詳しく説明すると、本実施形態では、事前学習で、正解関数 f_r が属する関数族 F （図 4 には、関数族 F の一部である関数 $f_1 \sim f_3$ のみを図示）が導出され、その後の再学習では、

ターゲットデータが良好に当てはまる関数（図4中の関数 f_v ）が選出される。ここで、再学習パラメータの数が1である場合、ターゲットデータが良好に当てはまる関数 f_v については、1自由度の範囲内で複雑な表現（例えば、非線形型の表現）が可能である。これは、再学習パラメータが推定モデルの入力側にあり、その下流側（より出力に近い側）には非線形型の演算モデル（具体的には、後述する推定結果出力モデル13）が存在しているためである。以上により結果、再学習パラメータの数を小さくしなければならない場合であっても、転移学習よりも高い精度にて物性値の実測値を推定する推定モデルが構築されるようになる。

[0068] 次に、本実施形態に係る推定モデルの構成について、図5を参照しながら説明する。本実施形態に係る推定モデル（以下、推定モデル10）は、図5に示すように、記述情報用演算モデル11と、識別情報用演算モデル12と、推定結果出力モデル13とによって構成されている。以下、各構成モデルについて説明する。

[0069] （記述情報用演算モデル）

記述情報用演算モデル11は、図5に示す通り、推定モデル10の入力側を構成する。この記述情報用演算モデル11は、化合物の構成に関する記述情報が入力されることで演算後記述情報を出力する。ここで、入力情報である記述情報は、化合物の構造を線形表記法に則って記述したグラフ表現である。グラフ表現とは、化合物を構成する原子の原子番号、及び化合物内の結合情報を列挙した可変長記述子のことである。

[0070] なお、記述情報については、グラフ表現に限定されず、例えば、化合物の構造についての文字列表現、具体的には、SMILES記法（Simplified Molecular Input Line Entry System）による記述子であってもよい。ただし、SMILES記述子を用いる場合には、同一化合物に対して多数の等価な表現方法が存在したり、環構造を持つ場合には隣接原子が文字列上の離れた場所に記載されるため長距離相関を持ったりするため、データ効率が悪い。一方、グラフ表現は、一意的に定まり、また、隣接原子の情報を明示的に機械

学習に与えることができるため、データ効率が良い。

また、記述情報の他の例としては、 k 次元（ k は自然数）の固定次元ベクトル表現であってもよい。ここで、固定次元ベクトル表現とは、化合物の特性のうち、安価に計算できるもの（原子数等）を事前に k 個定め、各化合物について、それらの特性を事前に計算してベクトル表現することで化合物の記述子としたものである。この固定次元ベクトル表現を用いる場合には、無数にある特性のうち、どの特性が物性値推定精度の向上に有効であるかを事前に検討する必要がある。一方、グラフ表現では、化合物のグラフから物性値推定に適した特性を機械学習にて自動的に特定することができるため、事前に人手によって有効な特性の選択を行う必要がない。

[0071] また、出力情報である演算後記述情報は、多次元の第一ベクトルである。つまり、記述情報用演算モデル11は、グラフ表現となった化合物の構造を示す記述情報を、固定次元ベクトルに変換する（エンコード）する演算モデルである。なお、演算後記述情報である第一ベクトルの次元数は、2以上の自然数であり、好適な値に設定される。また、第一ベクトルの次元数は、記述情報用演算モデル11中のパラメータの種類数に応じた数となっており、一般的には数十～数百に設定される。

[0072] また、記述情報用演算モデル11は、ニューラルネットワークによって構成されており、より具体的には、入力層、中間層及び出力層の各々が有する1又は複数のニューロン（ノード）が可塑性を持つ結合によって多数相互に結合された階層型ニューラルネットワークによって構成されている。なお、中間層の数、及び、各層におけるニューロンの数については、任意に設定することができる。また、記述情報用演算モデル11を構成するニューラルネットワークが、中間層を有さず、入力層及び出力層のみを有する二層構造のネットワークであってもよい。

[0073] また、記述情報用演算モデル11を構成するニューラルネットワークは、パラメータ ϕ を有する。このパラメータ ϕ は、機械学習（厳密には、事前学習）によって設定される学習パラメータであり、前述したパラメータ P に該

当する。また、パラメータ ϕ の数は、ニューラルネットワークの階層数及び各層のニューロンの数等に応じた数となる。なお、本実施形態におけるパラメータ ϕ の数については、事前学習に用いられる補助データの数に反映し、多数（例えば、数千～数万）設定されることが望ましい。これにより、記述情報用演算モデル 11 をなすニューラルネットワークの表現力が高められるようになる。

[0074] なお、本実施形態では、化合物のグラフ表現（具体的には、可変長記述子）が記述情報用演算モデル 11 に入力される。すなわち、本実施形態に係る記述情報用演算モデル 11 は、グラフニューラルネットワークであり、より厳密にはGoogle社考案のメッセージパッシングニューラルネットワーク（URL：<https://arxiv.org/pdf/1704.01212.pdf>にて公開された技術）である。メッセージパッシングニューラルネットワークは、化合物の構造に関するグラフ表現をニューラルネットワークで処理する際の代表的な構成である。

[0075] （識別情報用演算モデル）

識別情報用演算モデル 12 は、図 5 に示す通り、記述情報用演算モデル 11 と同様に、推定モデル 10 の入力側を構成する。この識別情報用演算モデル 12 は、条件の識別情報が入力されることで、演算後識別情報を出力する。ここで、入力情報である条件の識別情報とは、化合物の物性値を特定するために設定された条件を識別するための情報である。より詳しく説明すると、物性値を特定するために設定された条件は、ターゲットデータ（物性値の実測値）の取得条件、及び、補助データ（物性値の計算値）の取得条件であり、これら条件の各々に対して固有の情報が設定される。なお、本実施形態では、図 1 に示す「データ取得条件 i d」を示す番号が識別情報として設定される。ただし、これに限定されるものではなく、番号以外の情報、例えば番号以外の文字列又は記号等を識別情報として設定してもよい。

[0076] ちなみに、本実施形態において、データ取得条件 i d を示す識別情報は、one-hot 表現された上で識別情報用演算モデル 12 に入力される。また、本実施形態において、ターゲットデータの取得条件（すなわち、実測条

件)は、1種類のみとし、その条件に対して、データ取得条件 $i d = 0$ が識別情報として設定されることとする。他方、補助データの取得条件（すなわち、計算条件）は、数十種類存在し、それぞれの計算条件に対して、1以降の連番からなるデータ取得条件 $i d$ が識別情報として設定されることとする。

[0077] また、出力情報である演算後識別情報は、固有次元ベクトルとしての第二ベクトルである。つまり、識別情報用演算モデル 12 は、入力された識別情報を固定次元ベクトルに変換（エンコード）する演算モデルである。より詳しく説明すると、識別情報用演算モデル 12 は、図 6 に示すように、条件毎に設定された固有次元ベクトルによって構成された実数行列を含む演算モデルである。行列を構成する固有次元ベクトルの数（行数）は、条件の種類と同数である。そして、識別情報用演算モデル 12 に識別情報としてのデータ取得条件 $i d$ が入力されると、行列中、当該データ取得条件 $i d$ に対応する行の各成分値を抽出し、抽出した成分値を要素とする固有次元ベクトルを第二ベクトルとして出力する。一例を挙げて説明すると、データ取得条件 $i d = 2$ である場合、そのデータ取得条件 $i d$ が識別情報用演算モデル 12 に入力されると、データ取得条件 $i d = 2$ に対応する行（つまり、図 6 中、太枠で囲まれた行）の成分値を抽出し、各成分値からなる第二ベクトル（ $-0.3, 0.3, 0.2, -0.4$ ）を出力する。

[0078] ここで、識別情報用演算モデル 12 をなす行列の各成分値は、識別情報用演算モデル 12 中のパラメータ θ に相当する。このパラメータ θ は、事前学習によって設定される学習パラメータであり、前述のパラメータ Q に相当し、再学習によって更新される。また、パラメータ θ の数は、機械学習に用いられるターゲットデータ及び補助データの数に応じて設定される。より詳しく説明すると、識別情報用演算モデル 12 をなす行列の行数は、既知のターゲットデータの取得条件（すなわち、実測条件）の種類数と、既知の補助データの取得条件（すなわち、計算条件）の種類数とに基づいて設定される。

[0079] 一方、上記行列の列数は、比較的小さい値（例えば、数個程度）に設定さ

れる。これは、再学習に用いられる既知のターゲットデータの数が少ない場合にパラメータ数を多く設定すると過学習が生じる可能性があることを考慮し、少量のデータであっても過学習を回避して未測定の物性値を精度よく推定することが可能な推定モデルを構築するためである。

[0080] ここで、上記行列の列数は、識別情報用演算モデル 1 2 から出力される第二ベクトルの次元数に相当する。そして、本実施形態において、第二ベクトルは、比較的低次元のベクトルとなり、その次元数は、識別情報用演算モデル 1 2 のパラメータの数と同数であり、記述情報用演算モデル 1 1 のパラメータの数よりも小さく、且つ、推定結果出力演算モデル 1 3 のパラメータの数よりも小さくなっている。また、上述した理由から、第二ベクトルの次元数は、再学習で用いられる既知のターゲットデータの個数（すなわち、既知である物性値の実測値の数）よりも小さくなっている。

[0081] なお、本実施形態では、識別情報用演算モデル 1 2 が実数行列を用いた演算モデルであり、前述したように、入力された識別情報（データ取得条件 i d）に対応する行の各成分値を抽出し、その成分値からなる固有次元ベクトルを出力する。ただし、これに限定されるものではなく、入力された識別情報を行列計算以外の演算で所定次元数のベクトルに変換する関数を識別情報用演算モデル 1 2 として用いてもよい。

[0082] （推定結果出力モデル）

推定結果出力モデル 1 3 は、図 5 に示す通り、推定モデル 1 0 の出力側を構成する。この推定結果出力モデル 1 3 は、非線形型の演算モデルであり、化合物の構成に関する記述情報から得られた演算後記述情報、及び条件の識別情報から得られた演算後識別情報が入力されることで物性値の推定結果を出力する。

[0083] より厳密に説明すると、前述したように、記述情報用演算モデル 1 1 から演算後記述情報として第一ベクトルが出力され、識別情報用演算モデル 1 2 からは演算後識別情報として第二ベクトルが出力される。そして、推定結果出力モデル 1 3 は、第一ベクトル及び第二ベクトルから物性値（特徴）の

推定結果を出力する。厳密には、推定結果出力モデル 13 には、第一ベクトル及び第二ベクトルを結合させた結合ベクトルが入力される。推定結果出力モデル 13 は、入力された結合ベクトルから物性値の推定結果を出力する。ここで、結合ベクトルとは、第一ベクトルと第二ベクトルとを直列的に結合したベクトルであり、その次元数は、第一ベクトルの次元数と第二ベクトルの次元数との和に等しい。

[0084] また、推定結果出力モデル 13 は、ニューラルネットワークによって構成されており、より具体的には、入力層、中間層及び出力層からなる階層型ニューラルネットワーク（厳密には、広く一般的に用いられる全結合型ニューラルネットワーク）によって構成されている。なお、中間層の数、及び各層におけるニューロンの数については、任意に設定することができる。また、推定結果出力モデル 13 を構成するニューラルネットワークが中間層を有せず、入力層及び出力層のみを有する二層構造のネットワークであってもよい。

[0085] また、推定結果出力モデル 13 を構成するニューラルネットワークは、パラメータ ϕ を有する。このパラメータ ϕ は、機械学習（厳密には、事前学習）によって設定される学習パラメータであり、前述したパラメータ P に相当する。また、パラメータ ϕ の数は、ニューラルネットワークの階層数及び各層のニューロンの数等に応じた数となる。なお、本実施形態におけるパラメータ ϕ の数については、事前学習に用いられる補助データの数に反映し、多数（例えば、数千～数万）設定されることが望ましい。これにより、推定結果出力モデル 13 をなすニューラルネットワークの表現力が高められるようになる。

[0086] <本実施形態に係る特徴推定装置の構成>

次に、本実施形態に係る特徴推定装置 20 の構成例について、図 7 及び図 8 を参照しながら説明する。なお、図 7 では、外部インタフェースを「外部 I/F」と記載している。

[0087] 本実施形態に係る特徴推定装置 20 は、図 7 に示されるとおり、制御部 2

1、記憶部22、外部インタフェース23、入力装置24、出力装置25、及びドライブ26が電氣的に接続されたコンピュータである。ここで、特徴推定装置20をなすコンピュータは、1台であってもよく、あるいは複数台であってもよい。

[0088] 制御部21は、ハードウェアプロセッサであるCentral Processing Unit (CPU)、Random Access Memory (RAM)、Read Only Memory (ROM)等を含み、プログラム及びデータに基づいて各種のデータ処理を実行するように構成されている。なお、制御部21は、1又は複数の専用のハードウェアプロセッサを含んでもよい。

[0089] 記憶部22は、制御部21で実行される特徴推定用のプログラム31、及び、ニューラルネットワークの機械学習に利用するデータ群32等を記憶する。特徴推定用のプログラム31は、本実施形態に係る特徴推定方法によって機械学習を実施して推定モデル10を構築する処理をコンピュータに実行させるためのプログラムであり、本実施形態では記録媒体27から読み取られることで記憶部22に記憶される。機械学習用のデータ群32は、各化合物の構成に関する記述情報、既知であるターゲットデータ及び補助データ、並びにこれらのデータの取得条件の識別情報等である。

[0090] 外部インタフェース23は、外部装置と接続するためのインタフェースである。特徴推定装置20は、外部インタフェース23を介して、外部の機器（例えば、インターネットに接続された他のコンピュータ、及び、物性値の計測機器等）と通信することで、機械学習用のデータ群32の一部又は全部を取得することができる。

[0091] 入力装置24は、例えばマウス及びキーボード等からなり、ユーザの入力操作を受け付ける。特徴推定装置20は、ユーザが入力装置24を通じて物性値の実測値又は計算値等を入力することで、機械学習用のデータ群32の一部又は全部を取得することができる。

出力装置25は、例えばディスプレイ及びスピーカ等からなり、特徴推定装置20が推定した化合物の物性値等を出力するための装置である。

- [0092] ドライブ26は、例えば光学ディスクドライブ、磁気ディスクドライブ、ハードディスクドライブ及びメモリドライブ等からなり、記録媒体27に記憶されたプログラムを読み込むためのドライブ装置である。なお、ドライブ26の種類は、記録媒体27の種類に応じて適宜選択されてよい。また、記録媒体27に機械学習用のデータ群32の一部又は全部が記憶されている場合、特徴推定装置20は、ドライブ26が記録媒体27を読み取ることで上記機械学習用のデータ群32の一部又は全部を取得することができる。
- [0093] 記録媒体27は、コンピュータ及びその他装置又は機械等によって読み取り可能な各種のデータ及びプログラム等を、電氣的、磁氣的、光学的、機械的又は化学的作用によって蓄積する媒体である。本実施形態では、前述したように、記録媒体27に特徴推定用の学習プログラム31が記憶されている。なお、記録媒体27の一例として、CD及びDVD等の光学ディスク、磁気ディスク、ハードディスク等の内蔵ストレージ、フラッシュメモリ等の半導体メモリ、並びにメモリーカード等を挙げることができる。
- [0094] なお、特徴推定装置20の具体的なハードウェア構成に関しては、実施形態に応じて適宜、構成機器の追加、省略及び置換が可能である。
- [0095] 特徴推定装置20の構成を機能面から改めて説明すると、特徴推定装置20は、図8に示すように、データ記憶部41、学習データセット作成部42、モデル構築部43及び物性値推定部44を有する。これらの機能部は、前述した特徴推定装置20を構成するハードウェア機器と、ソフトウェアである特徴推定用のプログラム31とが互いに協働することで実現される。
- [0096] データ記憶部41は、機械学習用のデータ群32を記憶する。学習データセット作成部42は、データ記憶部41に記憶されたデータ群32から学習データセットを作成し、より詳しくは、事前学習に際して事前学習データセットを作成し、再学習に際して再学習データセットを作成する。
- [0097] モデル構築部43は、学習データセット（すなわち、化合物の構成に関する記述情報、データ取得条件の識別情報、及び化合物の物性値のセット）を用いた機械学習を実施して推定モデル10を構築する。本実施形態において

、モデル構築部43は、前述したように2段階のフェーズで機械学習を実施し、より詳しくは事前学習を実施した後に再学習を実施する。また、モデル構築部43は、それぞれの学習をニューラルネットワークによって実施する。

物性値推定部44は、モデル構築部43によって構築された推定モデル10を用いて、未測定である化合物の物性値の実測値を推定し、その推定結果を出力する。

[0098] なお、上述した特徴推定装置20の各機能部（すなわち、データ記憶部41、学習データセット作成部42、モデル構築部43及び物性値推定部44）は、のハードウェア構成は、専用のハードウェアであってもよいし、プログラムを実行する各種のプロセッサまたはコンピュータであってもよい。

また、各種のプロセッサには、ソフトウェア（プログラム）を実行して各機能部として機能する汎用的なプロセッサであるCPU（Central Processing Unit）、FPGA（Field Programmable Gate Array）等の製造後に回路構成を変更可能なプロセッサであるプログラマブルロジックデバイス（Programmable Logic Device：PLD）、及びASIC（Application Specific Integrated Circuit）等の特定の処理をさせるために専用に設計された回路構成を有するプロセッサである専用電気回路等が含まれる。

また、一つの機能部を、これら各種のプロセッサのうちの一つで構成してもよいし、同種または異種の二つ以上のプロセッサの組み合わせ、例えば、複数のFPGAの組み合わせ、または、FPGA及びCPUの組み合わせ等によって構成してもよい。また、複数の機能部を、各種のプロセッサのうちの一つで構成してもよいし、複数の機能部のうちの二つ以上をまとめて一つのプロセッサを用いて構成してもよい。

また、例えば、サーバおよびクライアント等のコンピュータに代表されるように、一つ以上のCPUとソフトウェアの組み合わせで一つのプロセッサを構成し、このプロセッサが複数の機能部として機能するケースが考えられる。ちなみに、本実施形態は、このケースに該当する。

また、システムオンチップ（System on Chip：S o C）等に代表されるように、複数の機能部を含むシステム全体の機能を一つの I C（Integrated Circuit）チップで実現するプロセッサを使用する形態がある。

さらに、上記各種のプロセッサのハードウェア構成は、より具体的には、半導体素子などの回路素子を組み合わせた電気回路（Circuitry）である。

[0099] <本実施形態に係る特徴推定方法について>

次に、本実施形態に係る特徴推定方法について説明する。なお、以下では、本実施形態に係る特徴推定方法を採用したデータ処理（以下、「特徴推定フロー」と言う。）の手順について、図 9 を参照しながら説明することとする。なお、図 9 に示す各ステップ S 0 0 1 ～ S 0 0 5 は、本実施形態に係る特徴推定方法の構成要素に相当する。

[0100] 特徴推定装置 2 0 を構成するコンピュータ（以下、単に「コンピュータ」と言う。）のユーザがフロー開始操作（例えば、特徴推定用のプログラム 3 1 を起動するための操作）を行い、特徴推定用のプログラム 3 1 が起動すると、これをトリガーとして特徴推定フローが開始される。

[0101] なお、特徴推定フロー開始前の時点で機械学習用のデータ群 3 2 が既已取得済みであって予めコンピュータの記憶部 2 2 に記憶されていてもよく、あるいは、特徴推定フローの開始後に上記のデータ群 3 2 を取得してもよい。以下では、特徴推定フローの開始時点で上記のデータ群 3 2 が記憶部 2 2 に記憶されているケースを例に挙げて説明することとする。

[0102] 特徴推定フローでは、まず、コンピュータが、記憶部 2 2 に記憶されたデータ群 3 2 から事前学習データセットを作成する（S 0 0 1）。本ステップ S 0 0 1 にて事前学習データセットを作成するにあたり、コンピュータは、化合物の物性値を推定する上で必要な既知の補助データを記憶部 2 2 から読み出す。この際、コンピュータは、補助データが取得された化合物の構成に関する記述情報（具体的には、グラフ表現）と、補助データの取得条件の識別情報（具体的には、計算条件に該当するデータ取得条件 i d）と、を併せて記憶部 2 2 から読み出す。そして、コンピュータは、補助データが示す物

性値の計算値、記述情報及び識別情報を 1 組とする事前学習データセットを、補助データの数だけ作成する。

[0103] なお、本実施形態では、数百～数千種類の化合物の各々について数十種類の計算条件の下で物性値が計算されるため、補助データソースが膨大なデータ数となる。このため、ステップ S 0 0 1 では、数千～数万個の事前学習データが作成されることになる。

[0104] 次に、コンピュータは、多数ある事前学習データセットを用いて事前学習を実施する (S 0 0 2)。つまり、本ステップ S 0 0 2 では、各補助データが得られた化合物の構成に関する記述情報と、当該各補助データの取得条件の識別情報と、当該各補助データが示す物性値の計算値とを用いて事前学習が実施される。この事前学習により、正解関数が属する関数族 F を多量の補助データソースから導出する。

[0105] より詳しく説明すると、事前学習を実施することにより、記述情報用演算モデル 1 1、識別情報用演算モデル 1 2、及び推定結果出力モデル 1 3 とが構築される。換言すると、事前学習では、コンピュータが上記 3 つの演算モデルの各々についてパラメータ θ 、 ϕ 、 ψ を設定する。

[0106] また、ステップ S 0 0 2 において、コンピュータは、ニューラルネットワークによって事前学習を実施する。具体的に説明すると、コンピュータは、事前学習において、各演算モデルのパラメータ θ 、 ϕ 、 ψ を初期化した後、予測誤差評価関数 L が最小化するように、それぞれのパラメータを、下記の式 (b 1) ～ (b 4) に従って連続的に計算する。

$$L(x, id; \theta_t, \phi_t, \psi_t) = [f(x, id; \theta_t, \phi_t, \psi_t) - y]^2 \quad (b 1)$$

$$\theta_{t+1} = \theta_t + \varepsilon \nabla_{\theta} L(x, \theta_t, \phi_t, \psi_t) \quad (b 2)$$

$$\phi_{t+1} = \phi_t + \varepsilon \nabla_{\phi} L(x, \theta_t, \phi_t, \psi_t) \quad (b 3)$$

$$\psi_{t+1} = \psi_t + \varepsilon \nabla_{\psi} L(x, \theta_t, \phi_t, \psi_t) \quad (b 4)$$

[0107] なお、上記式 (b 1) ～ (b 4) の各変数の定義は、以下の通りである。

L : 物性値の予測誤差評価関数

f : 推定モデルによる物性値の推定値

x : 化合物の構成に関する記述情報（例えば、グラフ表現）

id : 条件の識別情報（例えば、データ取得条件 id ）

y : 学習用のデータが示す物性値

θ : 記述情報用演算モデル 1 1 のパラメータ

ϕ : 識別情報用演算モデル 1 2 のパラメータ

ψ : 推定結果出力モデル 1 3 のパラメータ

ε : 学習係数

t : 学習ステップを示す回数（ $t = 1, 2, 3, \dots$ ）

[0108] そして、予測誤差評価関数 L の値が収束した時点で事前学習を終了する。

なお、予測誤差評価関数 L の勾配の推定には、広く一般的に用いられる誤差逆伝播法（バックプロパゲーション）が利用可能である。ただし、これに限定されるものではなく、誤差逆伝播法とそれ以外の方法（例えば勾配降下法等）を組み合わせ利用してもよい。

[0109] また、事前学習が終了した時点では、事前学習に用いられた補助データのそれぞれについて、補助データの取得条件の識別情報（つまり、計算条件に該当するデータ取得条件 id ）と対応する第二ベクトルが出力される。なお、これらの第二ベクトルを2次元平面に射影することにより、図10に示す散布図が得られる。

[0110] 次に、コンピュータは、記憶部22に記憶されたデータ群32から再学習データセットを作成する（S003）。本ステップS003にて再学習データセットを作成するにあたり、コンピュータは、化合物の物性値を推定する上で必要な既知のターゲットデータを記憶部22から読み出す。この際、コンピュータは、ターゲットデータが取得された化合物の構成に関する記述情報（具体的には、グラフ表現）と、ターゲットデータの取得条件の識別情報（具体的には、実測条件に該当するデータ取得条件 id ）と、を併せて記憶部22から読み出す。そして、コンピュータは、ターゲットデータが示す物性値の実測値、記述情報及び識別情報を1組とする再学習データセットを、ターゲットデータの数だけ作成する。

- [0111] なお、本実施形態では、数種類の化合物のみについて物性値が実測されるため、ターゲットデータソースは、少量のデータ数となる。このため、ステップS003では、数個～数十個の再学習データが作成されることになる。
- [0112] 次に、コンピュータは、少量の再学習データセットを用いて再学習を実施する（S004）。つまり、本ステップS002では、各ターゲットデータが得られた化合物の構成に関する記述情報と、当該各ターゲットデータの取得条件の識別情報と、当該各ターゲットデータが示す物性値の実測値とを用いて再学習が実施される。この再学習により、事前学習で導出された関数族Fのうち、ターゲットデータが良好に当てはまる関数 f_v を少量のターゲットデータソースから見つけ出すことができる。
- [0113] より詳しく説明すると、再学習では、コンピュータが、記述情報用演算モデル11中のパラメータ ϕ 、及び推定結果出力モデル13中のパラメータ ϕ を固定したまま、識別情報用演算モデル12中のパラメータ θ を、再学習データセット（すなわち、物性値の実測条件及びその実測値）に基づいて更新される。さらに詳しく説明すると、再学習では、識別情報用演算モデル12のパラメータ θ として、識別情報用演算モデル12をなす行列中の成分値を更新する。より厳密には、上記の行列中、ターゲットデータ取得条件の識別情報、すなわち、データ取得条件 $i d = 0$ と対応する行（図6に示すケースでは、1行目）の各成分値を更新する。
- [0114] また、ステップS004において、コンピュータは、ニューラルネットワークによって再学習を実施する。具体的に説明すると、コンピュータは、再学習において、識別情報用演算モデル12中のパラメータ θ のうち、更新対象であるデータ取得条件 $i d = 0$ と対応する行の各成分値を初期化した後、前述した（b3）に従って連続的に計算する。そして、予測誤差評価関数Lの値が収束した時点で再学習を終了する。なお、予測誤差評価関数Lの勾配の推定には、誤差逆伝播法（バックプロパゲーション）が利用可能であるが、それ以外の方法（例えば勾配降下法等）と組み合わせて利用してもよい。
- [0115] また、再学習が終了した時点では、再学習に用いられたターゲットデータ

について、当該ターゲットデータの取得条件の識別情報（すなわち、データ取得条件 $id = 0$ ）と対応する第二ベクトルが追加して出力される。この追加された第二ベクトルは、2次元平面に射影されることで、図11に示す散布図として、事前学習時に出力された第二ベクトルと共に可視化することができる。

[0116] そして、再学習が実施されることにより、未測定である化合物の物性値を推定する推定モデル10が構築される。この推定モデル10は、物性値が未測定である化合物の構成に関する記述情報と、物性値の実測条件の識別情報とを入力として、当該実測条件の下で特定される物性値の実測値を推定し、その推定結果を出力する。この推定モデル10によれば、未測定である化合物の物性値を精度よく推定（予測）することが可能となる。

[0117] 上記の推定モデル10について補足しておく、推定モデル10は、ターゲットデータが良好に当てはまる関数 f_v を表現したものであり、この関数 f_v は、識別情報用演算モデル12が出力する第二ベクトルが変わると、当然異なる形になる。かかる観点から言うと、再学習によって構築される推定モデル10は、第二ベクトルを変数とし、当該第二ベクトルの次元数を自由度として有する関数族を定義していると考えられる。つまり、第二ベクトルを固定した場合、推定モデル10は、化合物の構成に関する記述情報を入力して、当該化合物の物性値の推定結果を出力する関数となる。また、事前学習の終了後に、第二ベクトルを各条件の識別情報（データ取得条件 id ）と対応するように調整すれば、推定モデル10が各条件の識別情報に応じた関数となる。他方、推定モデル10が、いずれの識別情報とも対応しない第二ベクトルによって定義される関数となるようにすることもでき、この場合には、推定モデル10によって推定される物性値が、何らかの手法による計算値、あるいは何らかの実測環境下で実測された実測値であるものと期待される。

[0118] 次に、コンピュータは、ステップS004にて構築された推定モデル10を用いて、未測定である化合物の物性値の実測値を推定する（S005）。

図 1 に示すケースを例に挙げて説明すると、例えば、物性値が実測されていない化合物 2 について物性値の実測値を推定する場合には、化合物 2 の構造に関するグラフ表現と、実測条件の識別情報であるデータ取得条件 $id = 0$ と、を推定モデル 10 に入力する。これにより、化合物 2 の物性値の実測値が推定され、その推定結果が出力装置 25 によって出力される。

[0119] 以上までのステップが完了した時点で、特徴推定フローが終了する。そして、上述した特徴推定フローにより、実測値が欠落している化合物の物性値を補完することができる。

[0120] また、上述した特徴推定フローによれば、実測値のデータ（すなわち、ターゲットデータ）の数が少ない場合であっても、化合物の物性値の実測値を精度よく推定することができる。この点において、本実施形態は、特許文献 1 に記載された物性値推定方法よりも有利である。

[0121] より具体的に説明すると、特許文献 1 に記載の物性値推定方法のように、ニューラルネットワークによる機械学習を利用して化合物の物性値の実測値を推定（予測）する場合には、通常、実測値のデータ（すなわち、ターゲットデータ）がより多く得られているほど、推定精度が高くなる。しかしながら、物性値の実測は、手間及びコストを要するため、すべての化合物を対象として物性値を実測することは困難である。

[0122] 一方で、物性値は、前述したように、複数種類の計算条件（計算手法）によって計算することができ、各化合物の物性値をそれぞれの計算条件によって計算すれば、計算値のデータ（すなわち、補助データ）を多数取得することができる。そこで、本発明では、多数ある補助データを用いて事前学習を実施し、未測定である化合物の物性値を表現することが可能な正解関数が属する関数族 F を導出する。その後、少量のターゲットデータを用いて再学習を実施し、上記の関数族 F の中から、ターゲットデータが良好に当てはまる関数 f_v を選び出す。

[0123] ここで、本実施形態では、再学習にて更新されるパラメータが、推定モデル 10 の入力側にある識別情報用演算モデル 12 中のパラメータであり、識

別情報用演算モデル 12 の下流側（出力側）には、非線形型の推定結果出力モデル 13 が存在する。この結果、更新されるパラメータの数が少なく、例えば、その数が仮に 1 であったとしても、転移学習のようにベースライン補正とはならず、上記関数 f_v を自由度 1 の範囲内で、更新後のパラメータを変数とする非線形型の関数として表現することが可能となる。

以上により、少量のターゲットデータであっても、化合物の物性値の実測値を精度よく推定することが可能な推定モデル 10 を構築することが可能となる。

[0124] なお、上述の特徴推定フローは、既知のターゲットデータが少数である場合に特に有効であるが、当然ながら、既知のターゲットデータが多数取得される場合にも利用可能である。

また、推定モデル 10 の利用目的は、当然ながら、未測定である化合物の物性値を推定する目的に限られず、例えば、モデルの推定精度を検証する等の理由から実測済みの化合物の物性値を推定するために推定モデル 10 を利用してもよい。

[0125] <本実施形態の推定モデルの精度について>

以下では、上述した特徴推定フローによって構築された推定モデルの推定精度について行った検証実験と、その検証結果を説明する。

（検証用データ）

検証実験を実施するにあたり、化合物の物性値として、光の吸収エネルギー（換言すると、吸収波長）を示す検証用データを用意した。より詳しく説明すると、3000種類の化合物の各々について、26種類の汎関数のそれぞれによってDFT（Density Functional Theory 密度汎関数法）計算を行った。これにより、吸収エネルギーの計算結果を示す検証用データを、78000件（ $= 26 \times 3000$ ）用意した。なお、DFT計算には、量子化学計算ソフトウェアGaussianを使用し、基底関数を6-31Gとした。また、計算対象となった3000種類の化合物は、有機分子の物性値データベースであるQM9からランダムに抽出した。計算結果の一部の抜粋を図

12に示す。図12の横軸は、化合物の種類と対応する番号（化合物番号）を表しており、縦軸は、吸収エネルギーの計算値（単位：eV）を表している。

[0126] （検証実験）

検証実験では、先ず、検証用データを対象として統計解析を実施し、計算条件（具体的には、汎関数の種類）の違いによる計算結果の平均偏差を求めた。より詳しく説明すると、10点サンプリングで平均値補正を行った場合の平均偏差と、正確な平均値補正を行った場合の平均偏差とを算出した。前者の平均偏差は、0.27 eVであり、後者の平均偏差は、0.18 eVであった。

[0127] 次に、上述した特徴推定フローの手順にて推定モデルを構築した。ここで、推定モデルのうち、ニューラルネットワークによって構成される部分は、4層の全結合型ニューラルネットワークであり、具体的にはグラフニューラルネットワーク（厳密には、メッセージパッシングニューラルネットワーク）とした。また、推定モデルのうち、識別情報用演算モデルから出力されるベクトル（第二ベクトル）の次元数を10次元に設定した。

[0128] そして、26種類の計算条件のうち、20種類の計算条件（第一条件に相当）で3000種類の化合物の各々について得られた検証用データ、すなわち、60000件分の検証用データすべてを用いて事前学習を実施した。その後、残り6種類の計算条件（第二条件に相当）で5～20種類の化合物の各々について得られた検証用データを用いて再学習を実施した。以上までの過程を経て、第二条件に相当する6種類の計算条件の下で得られる物性値の計算値を推定する推定モデルが構築される。なお、第二条件に相当する6種類の計算条件とは、汎関数の種類としての「VSXC」、「MN12L」、「tHCTHhyb」、「HISS bPBE」、「BMK」及び「X3LYP」である。

[0129] 推定モデルの構築後、第二条件に相当する6種類の計算条件で3000種類の化合物の各々について得られた検証用データ、すなわち、18000件の検証用データを用いて推定精度の評価を行った。具体的には、3000種

類の化合物の各々について、上記 6 種類の計算条件で吸収エネルギーを計算した場合の計算値を、推定モデルによって推定した。そして、それぞれの推定結果を対応する検証用データと比較し、それぞれの計算条件について平均絶対誤差を求めた。検証結果を図 13 に示す。なお、図 13 の横軸は、再学習時に用いたデータの数を表しており、縦軸は、平均絶対誤差（単位：eV）を表している。

[0130] 検証実験の結果について説明すると、図 13 に示すように、第二条件に相当する 6 種類の計算条件の各々について、10 件のデータを用いて再学習した場合には、平均絶対誤差が 0.13 ~ 0.22 eV となった。つまり、上記 6 種類の計算条件の各々について、再学習に用いるデータ数が少量（10 件程度）であっても、各計算条件での計算値を推定モデルによって精度よく推定し得ることが明らかとなった。

[0131] <その他の実施形態について>

以上までに本発明の一実施形態について説明してきたが、上述の実施形態は、あくまでも一例に過ぎず、他の実施形態も考えられる。

[0132] 具体的に説明すると、上記の実施形態では、推定モデル 10 の入力側に、化合物の構成に関する記述情報から固定次元ベクトルを出力する記述情報用演算モデル 11 が設けられていることとした。ただし、これに限定されるものではなく、例えば、記述情報自体が固定次元ベクトルである場合には、記述情報用演算モデル 11 を省略してもよい。この場合には、識別情報用演算モデル 12 と推定結果出力モデル 13 のみによって推定モデル 10 が構成される。そして、識別情報用演算モデル 12 から出力される固定次元ベクトル（第二ベクトル）と、記述情報としての固定次元ベクトルとを結合し、その結合ベクトルが推定結果出力モデル 13 へ入力されることになる。

[0133] また、上述の実施形態では、ターゲットデータの取得条件（第二条件）が、1 つのみであり、補助データの取得条件（第一条件）よりも少ないこととしたが、これに限定されるものではない。第二条件に相当する条件が複数あってもよく、さらに第一条件に相当する条件より多くてもよい。

[0134] また、上述の実施形態では、事前学習において、識別情報用演算モデル 1 2 のパラメータ θ を設定し、具体的には、識別情報用演算モデル 1 2 をなす行列の各成分値を設定することとした。また、上記の実施形態では、再学習において、識別情報用演算モデル 1 2 中のパラメータ θ の一部を更新し、未測定の実測値を推定する推定モデル 1 0 を構築することとした。ただし、これに限定されるものではなく、上述の実施形態にベイズ推定の考えを導入してもよい。

[0135] つまり、ベイズ推定により、あるデータ取得条件の下で特定される物性値のデータが何らかの（確率的な）生成過程を通して生成されたと考え、その生成過程をモデル化する。このモデルでは、潜在変数 z としての第二ベクトルが正規分布に従って生成される。そして、ある化合物 n の構造のグラフ表現を記述情報用演算モデル 1 1 でエンコードした第一ベクトル x_n と、潜在変数 z である第二ベクトルから、既知関数 f （具体的には、推定結果出力モデル 1 3 により表現される関数）を通して物性値のデータ y を生成する。モデル化の作業としては、まず、事前学習を実施し、上記の既知関数 f を特定し、且つ、各データ取得条件（厳密には、各計算条件）に対応する潜在変数 z （第二ベクトル）を推定する。その後、再学習を実施する。再学習では、通常のベイズ推定と同様、物性値の実測値 y が何らかの潜在変数 z から生成されたとみなし、潜在変数 z が取り得る値の確率分布（事後分布）を計算する。具体的には、物性値の実測値 y のデータ（ターゲットデータ）を用い、その実測条件と対応する潜在変数 z についての事後分布をマルコフ連鎖モンテカルロ法でサンプリング近似する。その後、事後分布を利用して、物性値の実測値 y についての予測分布を計算し、例えば、予測分布の平均値を実測条件下での物性値の推定値とする。以上のような手順により、物性値の実測値を推定してもよい。

[0136] また、上述の実施形態では、化合物の構成に関する記述情報、及び実測条件の識別情報を入力として、化合物の物性値の実測値を推定するケースを例に挙げて説明したが、本発明は、その他の特徴を推定する場合にも適用可能

である。例えば、「人の声」を対象とし、その声が意味する語句（表音文字）を特徴として推定するケースにおいても、本発明は適用され得る。

[0137] より詳しく説明すると、推定対象者及び複数の非推定対象者の各々の声の音声データを解析して、それぞれの声が意味する語句（表音文字）を特定する。ここで、音声データが「記述情報」であり、声を発する人が「条件」に該当し、その人の名前又は識別IDが「識別情報」に該当する。また、非推定対象者が「第一条件」に該当し、推定対象者が「第二条件」に該当する。また、推定対象者の音声データから特定した語句（表音文字）が「ターゲットデータ」に相当し、非推定対象者の音声データから特定した語句（表音文字）が「補助データ」に該当する。そして、上述の実施形態と同様に、補助データのすべてを用いて事前学習を実施した後に、ターゲットデータを用いて再学習を実施することで、推定モデルが構築される。この推定モデルによれば、対象者の音声データと、対象者の名前又は識別IDとを入力として、上記音声データの声が意味する語句（表音文字）を推定（判定）することができる。

[0138] また、本発明のその他の適用例としては、「様々な書体で書かれた文字」を対象とし、その文字が表す語句の意味（例えば、数値）を特徴として推定するケースにおいても、本発明は適用され得る。

[0139] より詳しく説明すると、様々な書体で書かれた文字の画像データを解析して、それぞれの文字が表す数値を特定する。ここで、画像データが「記述情報」であり、書体が「条件」に該当し、その書体の名前等が「識別情報」に該当する。また、書体のうちの特定の書体が「第二条件」に該当し、残りの書体が「第一条件」に該当する。また、特定の書体で書かれた文字の画像データから特定した数値が「ターゲットデータ」に相当し、それ以外の書体で書かれた文字の画像データから特定した数値が「補助データ」に該当する。そして、上述の実施形態と同様に、補助データのすべてを用いて事前学習を実施した後に、ターゲットデータを用いて再学習を実施することで、推定モデルが構築される。この推定モデルによれば、特定の書体で書かれた文字の

画像データと、特定の書体の名前とを入力として、当該画像データが示す文字によって表される数値を推定（判定）することができる。

符号の説明

- [0140] 1 0 推定モデル
- 1 1 記述情報用演算モデル
- 1 2 識別情報用演算モデル
- 1 3 推定結果出力モデル
- 2 0 特徴推定装置
- 2 1 制御部
- 2 2 記憶部
- 2 3 外部インタフェース
- 2 4 入力装置
- 2 5 出力装置
- 2 6 ドライブ
- 2 7 記録媒体
- 3 1 プログラム
- 3 2 データ群
- 4 1 データ記憶部
- 4 2 学習データセット作成部
- 4 3 モデル構築部
- 4 4 物性値推定部
- f v ターゲットデータが良好に当てはまる関数
- F 関数族

請求の範囲

- [請求項1] コンピュータにより、対象の特徴を推定する特徴推定方法であって、
- 、
- コンピュータが、前記対象の構成に関する記述情報、前記特徴を特定する条件に対して設定された識別情報、及び前記条件の下で特定された前記特徴を用いた機械学習を実施して、前記記述情報及び前記識別情報から前記特徴を推定する推定モデルを構築し、
- 前記条件は、複数の第一条件と、前記複数の第一条件とは異なる第二条件と、を含み、
- 前記機械学習は、
- 前記記述情報、前記複数の第一条件の各々の前記識別情報、及び、前記複数の第一条件の各々の下で特定された前記特徴を用いた事前学習と、
- 前記第二条件の前記識別情報、及び、前記第二条件の下で特定された前記特徴を用いた再学習と、を含み、
- 前記事前学習では、コンピュータが、前記複数の第一条件の各々の前記識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、前記記述情報又は前記記述情報から得られた情報と前記演算後識別情報とが入力されて前記特徴の推定結果を出力する非線形型の推定結果出力モデルと、を構築し、
- 前記再学習では、コンピュータが、前記推定結果出力モデル中のパラメータを固定したまま前記識別情報用演算モデル中のパラメータを前記第二条件の前記識別情報、及び、前記第二条件の下で特定された前記特徴に基づいて更新し、前記記述情報及び前記第二条件の前記識別情報から前記第二条件の下で特定される前記特徴を推定する前記推定モデルを構築することを特徴とする特徴推定方法。
- [請求項2] 前記事前学習では、コンピュータが、前記記述情報が入力されて演算後記述情報を出力する記述情報用演算モデルと、前記識別情報用演

算モデルと、前記演算後記述情報と前記演算後識別情報とが入力されて前記特徴の推定結果を出力する前記推定結果出力モデルと、を構築し、

前記再学習では、コンピュータが、前記記述情報用演算モデル中のパラメータ、及び、前記推定結果出力モデル中のパラメータを固定したまま前記識別情報用演算モデル中のパラメータを前記第二条件の前記識別情報、及び、前記第二条件の下で特定された前記特徴に基づいて更新し、前記記述情報及び前記第二条件の前記識別情報から前記第二条件の下で特定される前記特徴を推定する前記推定モデルを構築する、請求項 1 に記載の特徴推定方法。

[請求項3] 前記演算後記述情報は、多次元の第一ベクトルであり、
前記演算後識別情報は、第二ベクトルであり、
前記推定結果出力モデルは、前記第一ベクトル及び前記第二ベクトルから前記特徴の推定結果を出力するモデルである、請求項 2 に記載の特徴推定方法。

[請求項4] 前記第二ベクトルの次元数は、前記記述情報用演算モデル中のパラメータの数よりも小さく、且つ、前記推定結果出力モデル中のパラメータの数よりも小さい、請求項 3 に記載の特徴推定方法。

[請求項5] 前記第二ベクトルの次元数は、前記再学習で用いられる前記第二条件の下で特定された前記特徴の数よりも小さい、請求項 4 に記載の特徴推定方法。

[請求項6] 前記識別情報用演算モデルは、前記条件毎に設定された固定次元ベクトルによって構成される行列を含む演算モデルであり、
前記再学習では、前記識別情報用演算モデル中のパラメータとして、前記行列中の成分値を更新する、請求項 1 乃至請求項 5 のいずれか一項に記載の特徴推定方法。

[請求項7] 前記機械学習において、コンピュータは、ニューラルネットワークにより前記事前学習及び前記再学習を実施する、請求項 1 乃至請求項

6 のいずれか一項に記載の特徴推定方法。

[請求項8] 前記推定モデルは、前記対象の前記特徴としての化合物の物性値を推定するモデルであり、

前記記述情報は、前記化合物の構造を線形表記法に則って記述した情報であり、

前記条件は、前記化合物の前記物性値を特定するために設定された条件である、請求項 1 乃至請求項 7 のいずれか一項に記載の特徴推定方法。

[請求項9] 前記第一条件は、前記物性値を計算するために設定された計算条件であり、

前記第二条件は、前記物性値を実測するために設定された実測条件である、請求項 8 に記載の特徴推定方法。

[請求項10] プロセッサを備え、対象の特徴を推定する特徴推定装置であって、
前記プロセッサは、前記対象の構成に関する記述情報、前記特徴を特定する条件に対して設定された識別情報、及び前記条件の下で特定された前記特徴を用いた機械学習を実施して、前記記述情報及び前記識別情報から前記特徴を推定する推定モデルを構築し、

前記条件は、複数の第一条件と、前記複数の第一条件とは異なる第二条件と、を含み、

前記機械学習は、

前記記述情報、前記複数の第一条件の各々の前記識別情報、及び、前記複数の第一条件の各々の下で特定された前記特徴を用いた事前学習と、

前記第二条件の前記識別情報、及び、前記第二条件の下で特定された前記特徴を用いた再学習と、を含み、

前記事前学習では、前記プロセッサが、前記複数の第一条件の各々の前記識別情報が入力されて演算後識別情報を出力する識別情報用演算モデルと、前記記述情報又は前記記述情報から得られた情報と前記

演算後識別情報とが入力されて前記特徴の推定結果を出力する非線形型の推定値出力モデルと、を構築し、

前記再学習では、前記プロセッサが、前記推定値出力モデル中のパラメータを固定したまま前記識別情報用演算モデル中のパラメータを前記第二条件の前記識別情報、及び、前記第二条件の下で特定された前記特徴に基づいて更新し、前記記述情報及び前記第二条件の前記識別情報から前記第二条件の下で特定される前記特徴を推定する前記推定モデルを構築することを特徴とする特徴推定装置。

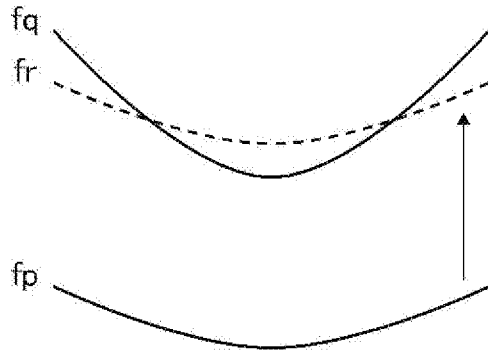
[請求項11] 請求項1乃至請求項9のいずれか一項に記載の特徴推定方法における前記機械学習を実施して前記推定モデルを構築する処理をコンピュータに実行させるためのプログラム。

[請求項12] 請求項1乃至請求項9のいずれか一項に記載の特徴推定方法における前記機械学習を実施して前記推定モデルを構築する処理をコンピュータに実行させるためのプログラムが記録された、コンピュータに読み取り可能な記録媒体。

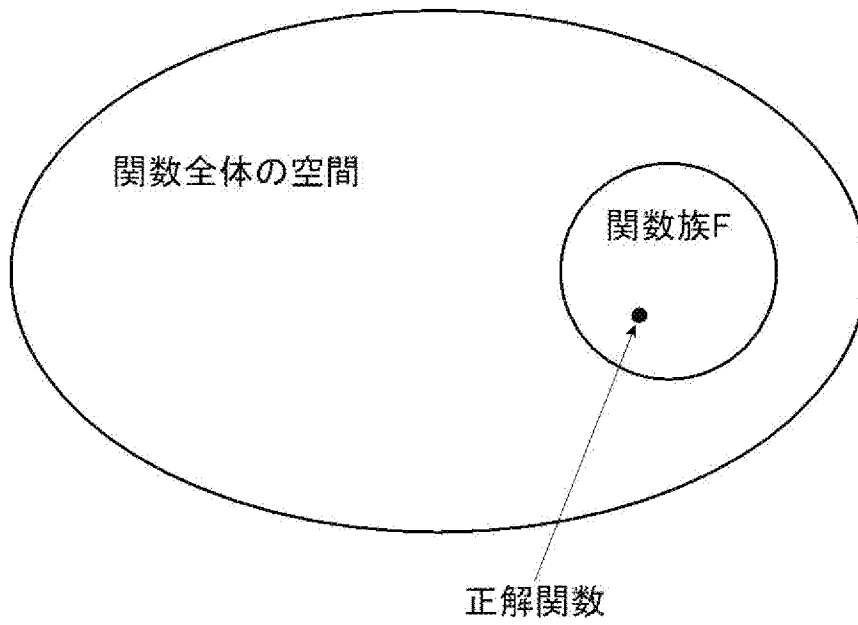
[図1]

データ取得条件id	データ取得条件	化合物1	化合物2	化合物3	化合物4	...	化合物N
0	実測A	1.9eV	欠落	3.9eV	欠落	...	2.4eV
1	計算B	2.4eV	3.6eV	4.6eV	2.1eV	...	2.6eV
2	計算C	3.1eV	2.5eV	4.0eV	1.3eV	...	3.1eV
3	計算D	2.8eV	2.3eV	4.8eV	2.5eV	...	3.2eV
⋮	⋮	⋮	⋮	⋮	⋮		⋮
K-1	計算Y	0.7eV	3.8eV	3.9eV	2.4eV	...	3.4eV
K	計算Z	1.1eV	2.2eV	4.0eV	1.9eV	...	2.2eV

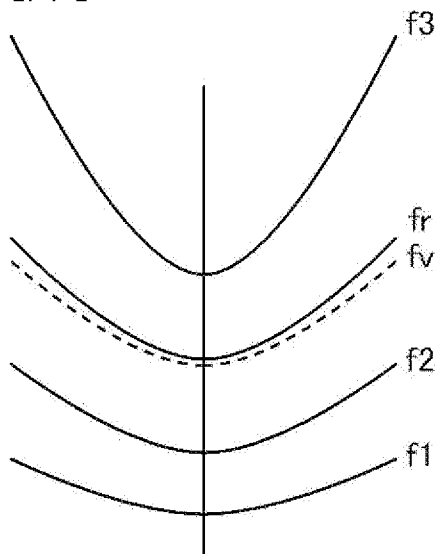
[図2]



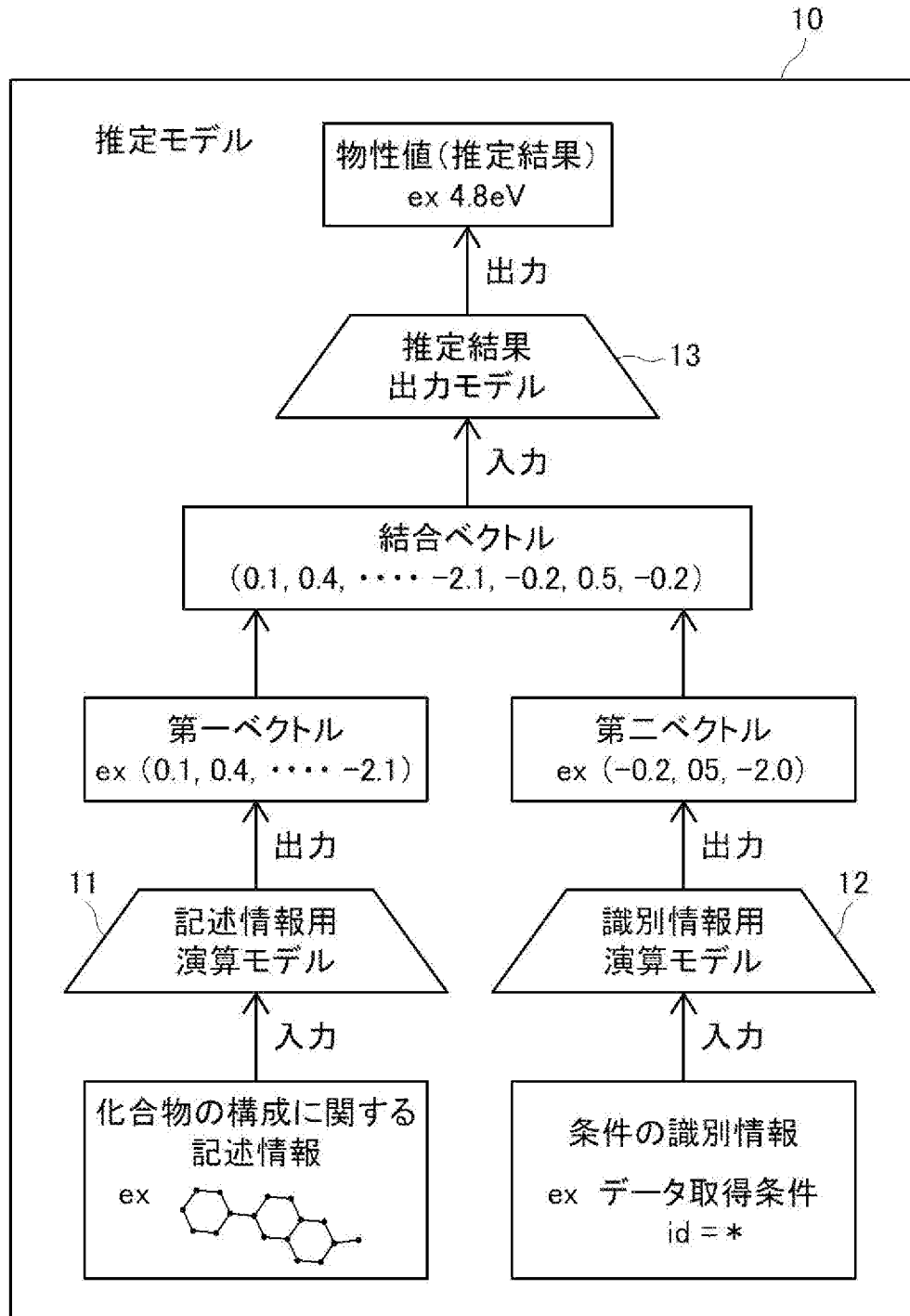
[図3]



[図4]



[図5]

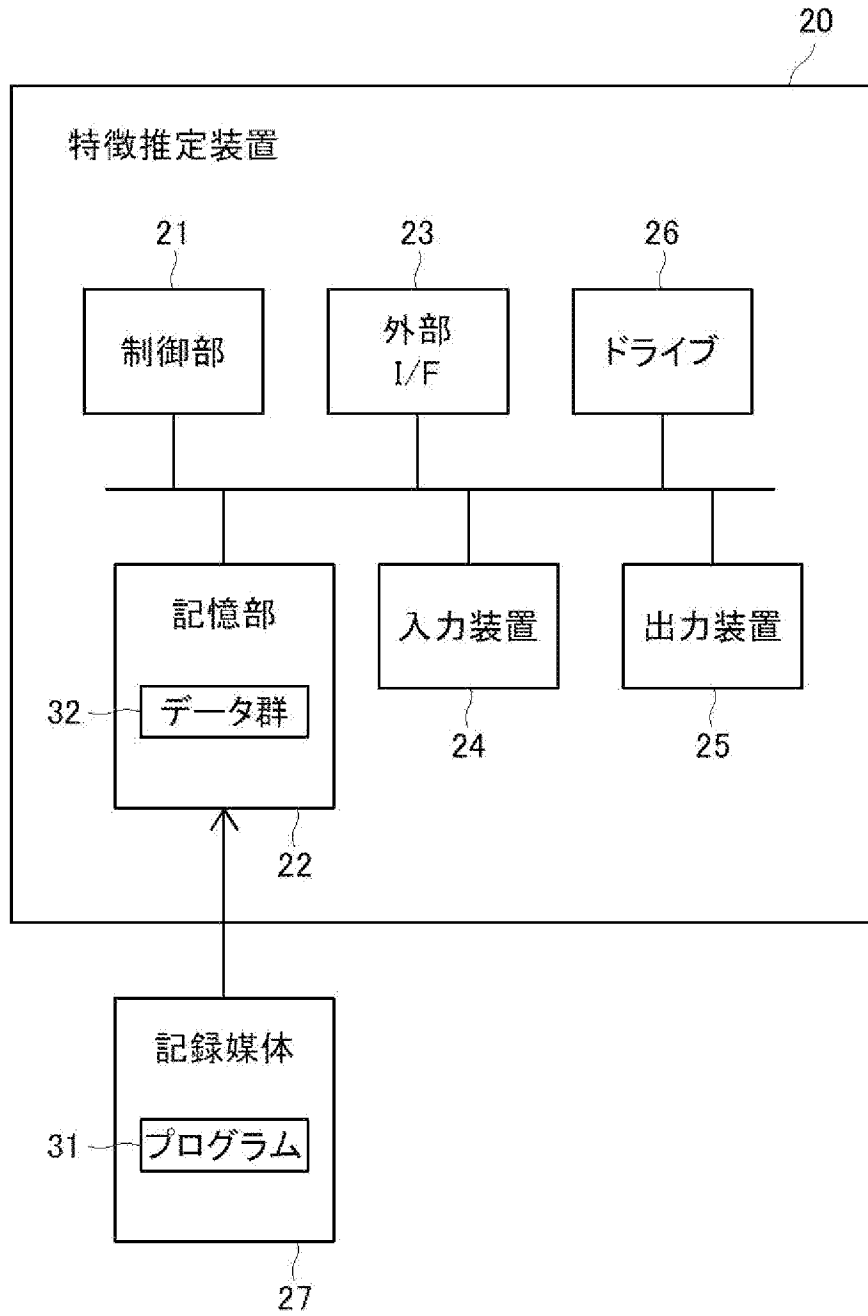


[図6]

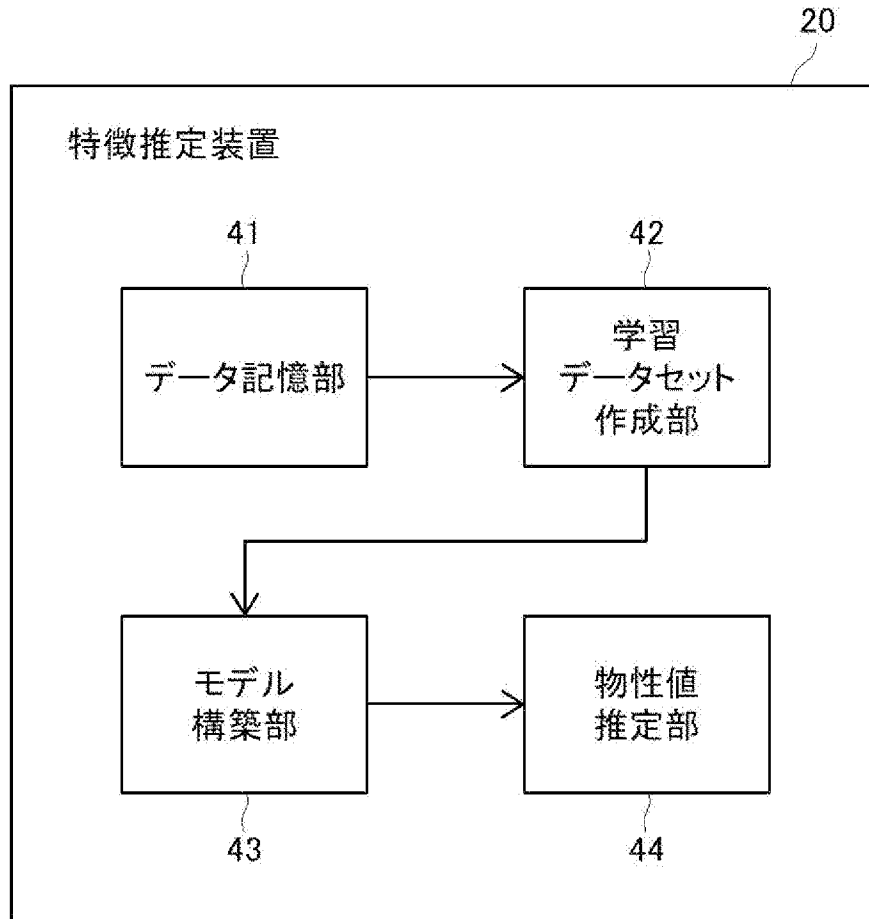
対応する
データ
取得条件id

0	0.0	0.4	0.1	0.1
1	-0.4	0.1	-0.1	-0.3
2	-0.3	0.3	0.2	-0.4
3	-0.2	0.1	0.4	0.3
⋮	⋮	⋮	⋮	⋮
K-2	0	0.3	0.1	0.3
K-1	-0.1	-0.3	-0.5	0.5
K	-0.1	0.1	-0.3	0.0

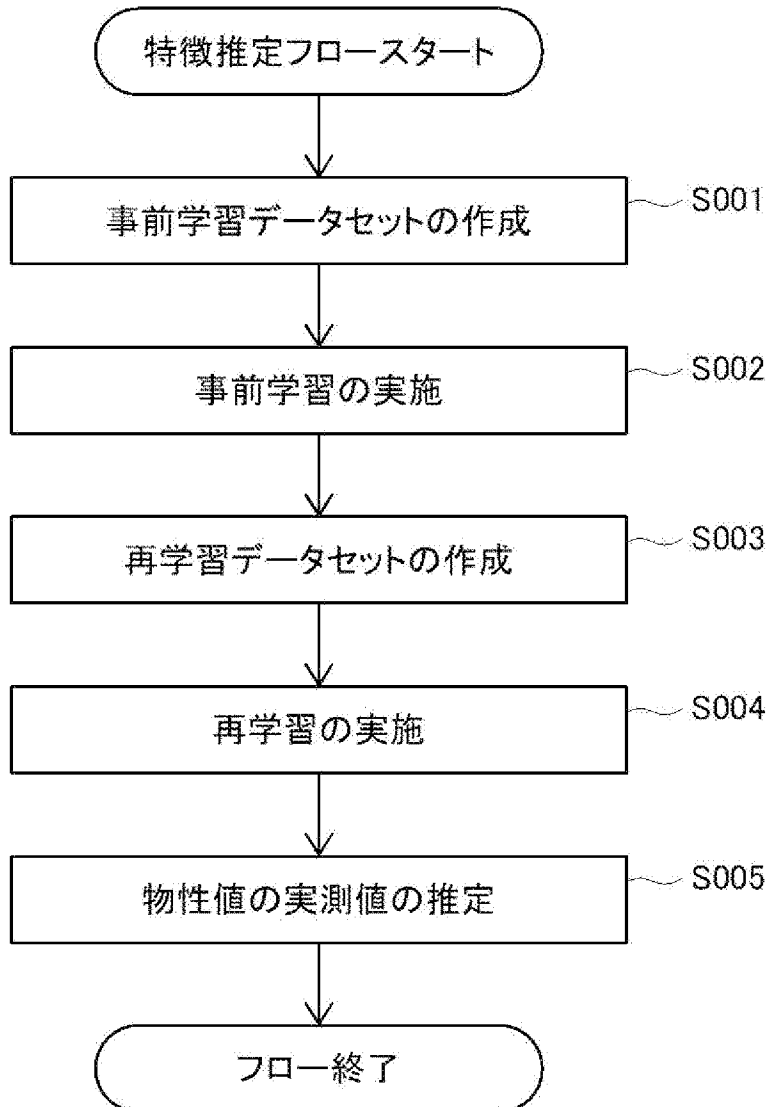
[図7]



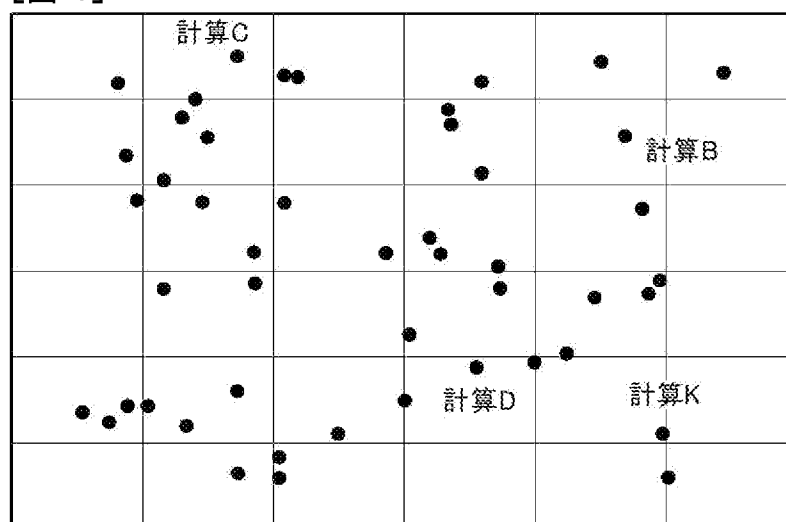
[図8]



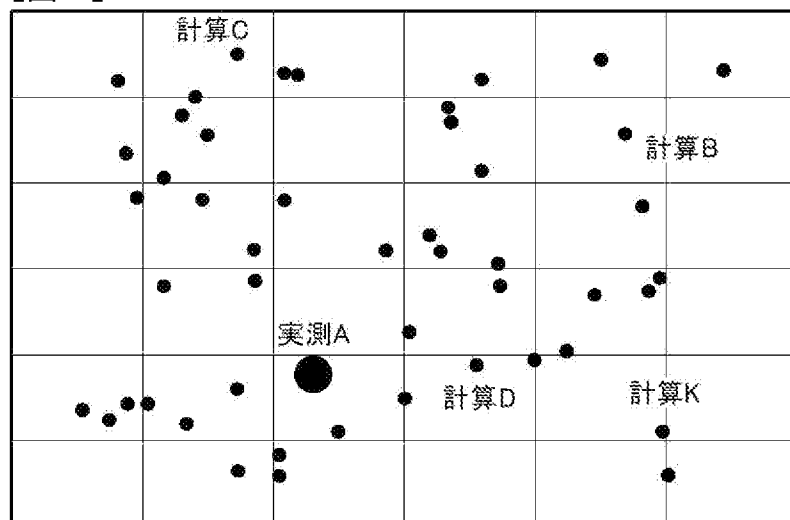
[図9]



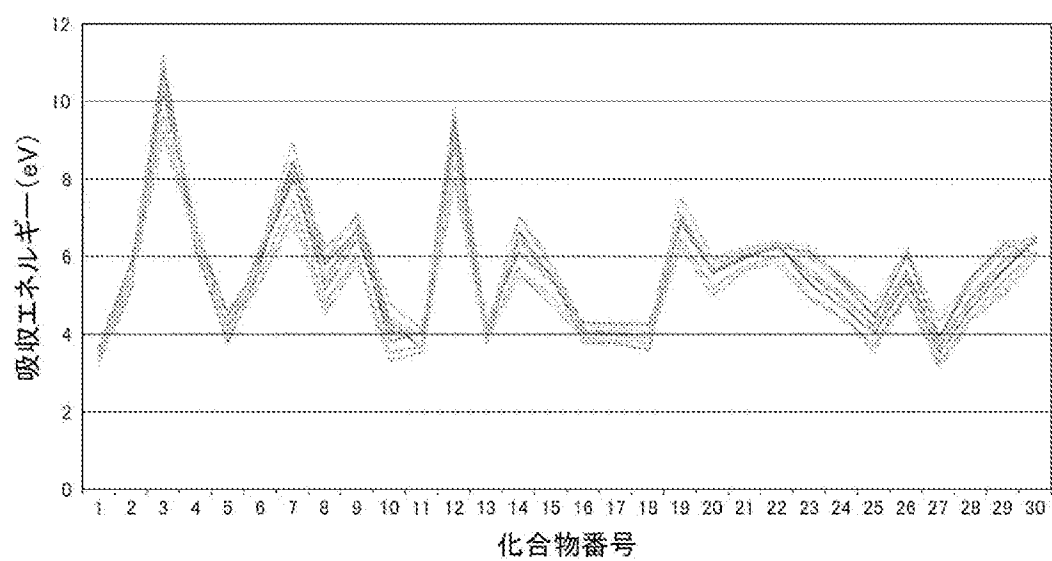
[図10]



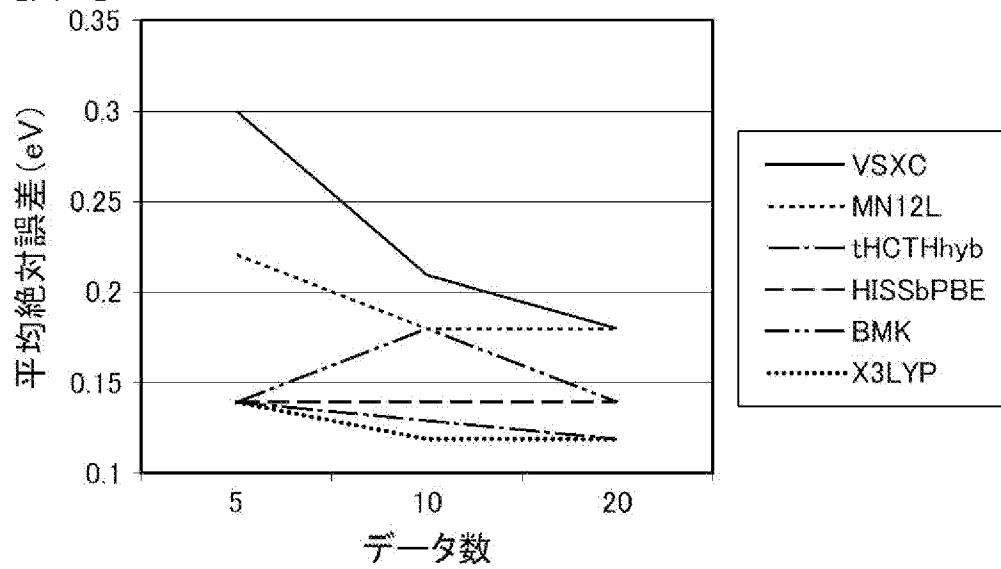
[図11]



[図12]



[図13]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/000837

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/00 (2019.01) i

FI: G06N20/00 130

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2012-226732 A (NIPPON STEEL CORP.) 15.11.2012 (2012-11-15) paragraphs [0006]-[0015]	1-12



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"I" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

18 March 2020 (18.03.2020)

Date of mailing of the international search report

31 March 2020 (31.03.2020)

Name and mailing address of the ISA/

Japan Patent Office

3-4-3, Kasumigaseki, Chiyoda-ku,

Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2020/000837

Patent Documents
referred in the
Report

Publication
Date

Patent Family

Publication
Date

JP 2012-226732 A

15 Nov. 2012

(Family: none)

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 20/00(2019.01)i FI: G06N20/00 130														
B. 調査を行った分野														
調査を行った最小限資料（国際特許分類（IPC）） G06N20/00														
最小限資料以外の資料で調査を行った分野に含まれるもの <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2020年	日本国実用新案登録公報	1996 - 2020年	日本国登録実用新案公報	1994 - 2020年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2020年													
日本国実用新案登録公報	1996 - 2020年													
日本国登録実用新案公報	1994 - 2020年													
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）														
C. 関連すると認められる文献														
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号												
A	JP 2012-226732 A（新日本製鐵株式会社）15.11.2012（2012 - 11 - 15） 段落[0006]-[0015]	1-12												
☐ C欄の続きにも文献が列挙されている。 ☒ パテントファミリーに関する別紙を参照。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの</td> </tr> <tr> <td>“A” 特に関連のある文献ではなく、一般的技術水準を示すもの</td> <td>“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの</td> </tr> <tr> <td>“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの</td> <td>“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの</td> </tr> <tr> <td>“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）</td> <td>“&” 同一パテントファミリー文献</td> </tr> <tr> <td>“O” 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの	“A” 特に関連のある文献ではなく、一般的技術水準を示すもの	“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの	“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの	“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	“&” 同一パテントファミリー文献	“O” 口頭による開示、使用、展示等に言及する文献		“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	
* 引用文献のカテゴリー	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの													
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの	“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの													
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの													
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	“&” 同一パテントファミリー文献													
“O” 口頭による開示、使用、展示等に言及する文献														
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献														
国際調査を完了した日 18.03.2020	国際調査報告の発送日 31.03.2020													
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 多胡 滋 5B 3562 電話番号 03-3581-1101 内線 3545													

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2020/000837

引用文献	公表日	パテントファミリー文献	公表日
JP 2012-226732 A	15.11.2012	(ファミリーなし)	